



universität
wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Development of a KNIME workflow for the retrieval of associations between orphan diseases and their possible drug repurposing candidates“

verfasst von / submitted by

Jana Gurinova

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Magistra der Pharmazie (Mag. Pharm.)

Wien, 2018 / Vienna, 2018

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

A 449

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Diplomstudium Pharmazie

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Gerhard Ecker

Mitbetreut von / Co-Supervisor:

Dipl.-Ing. (FH) Dr. Daniela Digles

DEVELOPMENT OF A KNIME WORKFLOW FOR THE RETRIEVAL OF
ASSOCIATIONS BETWEEN ORPHAN DISEASES AND THEIR
POSSIBLE REPURPOSING CANDIDATES

JANA GURINOVA



universität
wien

Department of Pharmaceutical Chemistry
Faculty of Life Sciences
University of Vienna

19th November 2018

JANA GURINOVA: *Development of a KNIME workflow for the retrieval of associations between orphan diseases and their possible repurposing candidates*, The workflow was developed using the KNIME® Analytics Platform, © 19th November 2018

SUPERVISORS:

Univ.-Prof. Mag. Dr. Gerhard F. Ecker

Dipl.-Ing. (FH) Dr. Daniela Digles

LOCATION:

Vienna, Austria

*In order to carry a positive action,
we must develop here a positive vision.*

— DALAI LAMA

*In a gentle way,
you can shake the world.*

— MAHATMA GANDHI

ACKNOWLEDGEMENTS

I would very much like to acknowledge first and foremost my co-supervisor and thank her for her help and guidance during my work on the diploma thesis. Her insight and suggestions were very valuable, and helped shape this work into the scientific contribution it is today.

Secondly I would like to acknowledge my professor Gerhard F. Ecker, and thank him for giving me this opportunity to work on such an interesting and challenging topic and for thusly starting my scientific journey.

For her help with the analysis of a subset of the workflow results I would like to thank Jerneja Kladnik, who manually curated the subset and by doing so helped to statistically characterise the results of the workflow and thus to validate it.

Furthermore I would like to acknowledge the Open PHACTS Foundation that makes a continued use of the Open PHACTS Discovery Platform possible. I would also like to acknowledge the support this work has received from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. [115191], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution of EFPIA companies.

On a more personal note, the two following pages are dedicated to two people who are an integral part of my life. They are dedicated to my partner Korbinian Groll, who is a very kind, patient and loving man, and to my mother Magdalena Resch-Gurin, who is a strong and courageous woman. You make my world a better place.

BARE NECESSITIES

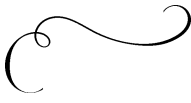


*Look for the bare necessities
The simple bare necessities
Forget about your worries and your strife
I mean the bare necessities
Are Mother Nature's recipes
That bring the bare necessities of life*

*Wherever I wander
Wherever I roam
I couldn't be fonder of my big home
The bees are buzzin' in the tree
To make some honey just for me
When you look under the rocks and plants
And take a glance at the fancy ants
Then maybe try a few*

*The bare necessities of life will come to you
They'll come to you!*





*"Hope" is the thing with feathers —
That perches in the soul —
And sings the tune without the words —
And never stops — at all —*

*And sweetest — in the Gale — is heard —
And sore must be the storm —
That could abash the little Bird
That kept so many warm —*

*I've heard it in the chilliest land —
And on the strangest Sea —
Yet, never, in Extremity,
It asked a crumb — of Me.*

— Emily Dickinson

*Nádej, tá má krídla —
Hniezdi niekde v duši —
Spiera tíško piesne bez slov —
I keď búrku tuší —*

*Sladký hlasok počuť naprieč —
Aj keď hromy divo bijú —
Neutíšia malé vtáča
Čo skrýja, nedobyjú —*

*Počuť ho aj v chladných krajoch —
Kde lahla mračien tóna —
Je útechou a nežiada
ani omrvinku — zo mňa.*



Seltene Erkrankungen betreffen weltweit über 400 Millionen Menschen. Sie werden über ihre Häufigkeit beziehungsweise ihre Seltenheit definiert, und daher zwangsläufig auch über eine geringe Zahl an Patienten pro Krankheit. Im Vergleich zur hohen Gesamtzahl an Patienten mangelt es jedoch an medikamentösen Behandlungen für die rund 8,000 bekannten seltenen Erkrankungen. In den USA sind diesbezüglich rund 500 Arzneimittel auf dem Markt und in der EU sind es rund 80. Neue Arzneimittel könnten hier Abhilfe schaffen, und eine Strategie um in möglichst kurzer Zeit Arzneimittel für neue Indikationen verfügbar zu machen ist die sogenannte Umnutzung oder Neupositionierung von Arzneimitteln (eng. drug repurposing oder drug repositioning), also die Verwendung von bekannten Arzneimitteln für Indikationen abseits der ursprünglichen Zulassung.

Die Neupositionierung von Arzneimitteln war Gegenstand zahlreicher wissenschaftlicher Publikationen des letzten Jahrzehnts. Der Fokus liegt bei diesen Publikationen üblicherweise auf der Neupositionierung für eine bestimmte Krankheit oder eine Gruppe von Krankheiten. Eine kleine Anzahl dieser Publikationen widmet sich *in silico* Methoden für eine übergreifende Neupositionierung über Krankheiten und Klassen hinweg. Um die hohe Anzahl an Patienten möglichst deckend mit Therapiemöglichkeiten versorgen zu können, liegt es nahe eine übergreifende Methode zur Neupositionierung zu verwenden. Die hier präsentierte Arbeit widmet sich dem Entwurf und der Ausführung einer solchen Methode.

Die hier beschriebene Methode ist eine Dreiecksmethode. Sie erstellt Verbindungen zwischen Krankheiten, deren Zielproteinen, und den Arzneistoffen welche Wirkungen bei diesen Zielproteinen aufweisen. Die Erweiterung der Verbindung zwischen einer Krankheit und einem Arzneistoff um das Zielprotein der Krankheit ermöglicht die Entdeckung bisher unbekannter Zusammenhänge zwischen Krankheiten und Arzneistoffen. Das Workflow-Management-System welches für die Ausführung dieser Analyse verwendet wurde ist KNIME®.

Der Workflow beruht auf der Zusammenlegung von drei Datenbanken. Kurz gesagt beginnt der Workflow mit einer Liste seltener Erkrankungen aus der Datenbank Orphanet, verbindet die Liste mit der Open PHACTS Discovery Plattform um Zielproteine abzurufen, und verknüpft diese Zielproteine zuletzt mit zugehörigen Arzneistoffen aus der Datenbank DrugBank. Bei dieser Herangehensweise wird die Tatsache genützt, dass ein Zielprotein bei mehreren Erkrankungen eine Rolle spielen kann, und dass ein Arzneimittel welches für dieses Zielprotein im Rahmen einer Erkrankung entwickelt wurde, unter Umständen auch für die Behandlung einer anderen Erkrankung herangezogen werden kann.

Diese Methode basiert ausschließlich auf Daten und die daraus resultierende Anreicherung von (für eine Neupositionierung sinnvollen) Verbindungen zwischen Krankheiten und Arzneimitteln im endgültigen Datensatz beträgt rund 60%. Dieser Prozentsatz wurde durch Untersuchungen von Teildatensätzen der

endgültigen Ergebnisse bestimmt. Diese Untersuchungen ergaben dass rund zwei aus drei Verbindungen entweder bereits publiziert waren, mittels manueller Überprüfung des Zusammenhangs als sinnvoll eingestuft werden konnten, klinische Studien diesbezüglich im Gang waren oder eine Verbindung bereits zugelassen war.

Das Ziel dieser Arbeit war es einen Workflow zu erstellen welcher Vorschläge zur Neupositionierung von Arzneimitteln für seltene Erkrankungen erbringen kann, und dieses Ziel wurde hier erreicht.

ABSTRACT

Worldwide approximately 400 million people are affected by orphan diseases. Orphan diseases are defined by their low prevalence and therefore a low number of patients per disease. Treatments for these conditions - of which there are around 8,000 - are however few and far between, amounting to around 500 drugs with a marketing approval for an orphan disease in the US and around 80 in the EU. The need for new therapies can therefore clearly be seen. One of the possibilities to meet the needs of the rare disease patient community might lie in drug repurposing, the application of approved drugs for the treatment of new indications. A significant advantage lies in the fact that drug repurposing can provide safe drugs in a fraction of the time it would take to develop a new drug.

Drug repurposing has been addressed in a number of publications in the last decade, most of them focusing on drug repurposing for specific diseases or disease classes, while a small number of them elaborates on *in silico* methods for comprehensive repurposing across many diseases. Given the high number of orphan diseases, the patient community could obviously profit from a general repurposing method being applied to all orphan diseases. The work described in this thesis is dedicated to the design and implementation of such a method.

The method described in this thesis is a triangulation between diseases, their targets and drugs that act on these targets. Such a triangulation is able to retrieve previously unknown associations between diseases and drugs by expanding the connection between a disease and its drug through the introduction of an intermediate step where target information is taken into account. KNIME® is the workflow management system that was used to execute this triangulation.

The workflow is based on the integration of three powerful databases. In a nutshell the workflow starts with a list of orphan diseases from the Orphanet database, which is then connected to the Open PHACTS Discovery Platform for the retrieval of targets, and the targets are subsequently connected to respective drugs from DrugBank. The idea behind this approach is that one target can be related to multiple diseases, and that a drug that has been developed for such a target in the context of one disease can potentially be used for the treatment of another disease.

The method relies exclusively on data mining and results in an enrichment of useful associations between diseases and drugs of approximately 60%. This means that in the inspected subsets of results roughly two out of three associations between a disease and a drug were found to be either mentioned in publications, determined to be useful by a manual check, under investigation in clinical trials or already approved.

The aim of this work, which was to create a workflow capable of proposing drug repurposing candidates for orphan diseases, was therefore achieved.

CONTENTS

List of Figures	xv
List of Tables	xvii
List of Listings	xviii
Acronyms	xix
I INTRODUCTION	1
1 ORPHAN DISEASES	3
1.1 Socio-economic aspects	6
1.2 State of the drug repurposing art	10
1.3 Validation strategies for computational drug repurposing	13
II MAIN BODY	15
2 METHODS	17
2.1 Databases	17
2.1.1 Orphanet	17
2.1.2 Open PHACTS Discovery Platform	19
2.1.3 DisGeNET	20
2.1.4 DrugBank	21
2.1.5 Rare Disease Repurposing Database	22
2.1.6 repoDB	26
2.1.7 EMBL-EBI webservices	28
2.2 KNIME® Analytics Platform	29
2.2.1 Data Format: XML	30
2.2.2 Data Retrieval: Download, API Calls	37
2.2.3 Mining of textual data	42
2.3 The Workflow	45
2.3.1 Extraction of diseases	45
2.3.2 Extraction of genes	47
2.3.3 Extraction of drugs	51
2.3.4 Text mining	55
2.3.5 Analysis of results	62
3 RESULTS	63
3.1 Workflow Results	63
3.2 Validation	71
3.2.1 Validation with the Rare Disease Repurposing Database	71
3.2.2 Validation with repoDB	72
4 DISCUSSION	73
4.1 Limitations	73
4.2 Possibilities for adaptation	74
4.2.1 Gene retrieval segment	74
4.2.2 Drug retrieval segment	76
4.2.3 Text mining segment	76

III	APPENDIX	77
A	POSTERS	79
A.1	Linking Life Science Data: Design to Implementation, and Beyond	79
A.2	The ISMC and the YMCS	82
A.3	255 th ACS National Meeting & Exposition	85
	BIBLIOGRAPHY	93

LIST OF FIGURES

Figure 1	Triangulation between diseases, targets and drugs . . .	17
Figure 2	Example of an Orphanet entry	18
Figure 3	DisGeNET RDF	20
Figure 4	Example of a DisGeNET entry	21
Figure 5	Example of a DrugBank entry	22
Figure 6	Basic KNIME® workflow	29
Figure 7	KNIME® workbench	30
Figure 8	KNIME® XML Reader and XPath nodes	30
Figure 9	XPath in KNIME®	32
Figure 10	Node repository - Import nodes	38
Figure 11	Node repository - XML nodes	38
Figure 12	XML Reader - configuration	39
Figure 13	OPS_Swagger - access keys	40
Figure 14	Disease URL - browser result	41
Figure 15	OPS_Swagger - API	42
Figure 16	Word cloud	43
Figure 17	KNIME® Textprocessing	43
Figure 18	Schematic view of the workflow	45
Figure 19	Nodes for the extraction of UMLS identifiers of orphan diseases	45
Figure 20	'Rare diseases and cross-referencing' dataset as seen in the KNIME® table	46
Figure 21	Results of the identifier extraction as seen in the KNIME® table	46
Figure 22	Nodes for the retrieval of genes through Open PHACTS based on UMLS identifiers and for the extraction of genes from the 'Rare diseases with their associated genes' dataset from Orphadata	47
Figure 23	Disease-gene associations from Orphanet before and after identifier mapping	51
Figure 24	Nodes for the retrieval of drugs based on genes through Open PHACTS and the addition of annotations directly from DrugBank	51
Figure 25	Gene-drug associations retrieved through Open PHACTS before and after an additional annotation with DrugBank data	52
Figure 26	A step by step extraction of annotations from the DrugBank XML as seen in the various KNIME® tables	53
Figure 27	Nodes for the retrieval of publication abstracts based on PubMed identifiers	55
Figure 28	Examples of XPath expressions used to search for terms within an XML file	56

Figure 29	KNIME [®] table resulting from a dictionary applied within an <i>XPath</i> node	56
Figure 30	Nodes used for a classification of abstracts after a dictionary-based text mining	57
Figure 31	Table resulting from a text mining approach based on a dictionary	57
Figure 32	Nodes for a vector-based text mining approach	58
Figure 33	Table as seen after the conversion of publication abstracts from the <i>string</i> into the <i>document</i> format	59
Figure 34	Terms as extracted by the <i>Bag of Words</i> node	60
Figure 35	Bitvector of terms within publication abstracts	60
Figure 36	Disease-gene association types as seen within the ‘Rare diseases and cross-referencing’ dataset	61
Figure 37	Nodes for the analysis of the workflow results and for the evaluation of drug repurposing candidates	62
Figure 38	Types of disease-gene associations with respective counts	64
Figure 39	Distribution of the number of diseases per associated number of genes	65
Figure 40	Distribution of the number of disease per associated number of genes for a maximum of 20 diseases and 300 genes	66
Figure 41	Distribution of unique disease-drug associations and their categories as assigned by text mining	67
Figure 42	Results from the sensitivity-based validation with the Rare Disease Repurposing Database	72
Figure 43	Results from the sensitivity- and specificity-based validation with the repoDB	72
Figure 44	Abstract for the Open PHACTS closing conference	80
Figure 45	Poster for the Open PHACTS closing conference	81
Figure 46	Abstract for the ISMC and YMCS	83
Figure 47	Poster for the ISMC and YMCS	84
Figure 48	Abstract for the 255 th ACS National Meeting & Exposition	85
Figure 49	Abstract for the CINF Scholarship for Scientific Excellence	86
Figure 50	Poster for the 255 th ACS National Meeting & Exposition	91

LIST OF TABLES

Table 1	Regulatory environments of orphan diseases	4
Table 2	BURQOL-RD’s ten rare diseases	7
Table 3	Incentives for the development of orphan drugs	8
Table 4	Drug repurposing examples for orphan diseases	13
Table 5	Excerpt from Table 1 in the Rare Disease Repurposing Database	24
Table 6	Excerpt from Table 2 in the Rare Disease Repurposing Database	25
Table 7	Example entries in the repoDB dataset	27
Table 8	Results at different positions within the workflow	68

LIST OF LISTINGS

Figure 1	Example XML	31
Figure 2	Orphanet XML: Full entry on cross-referencing of cystic fibrosis	36
Figure 3	DrugBank XML segment: Targets of Ivacaftor	37
Figure 4	Example XML resulting from an ‘Associations for Dis- ease’ API call	48
Figure 5	Python script for the mapping of UniProt and NCBI identifiers	50
Figure 6	Expressions used in the <i>Rule-based Row Splitter</i> to retain only correct combinations of disease-gene associations and the mechanisms of action of drugs	62

ACRONYMS

API	Application Programming Interface
BURQOL-RD	Social Economic Burden and Health-Related Quality of Life in Patients with Rare Diseases in Europe
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
FDA	U.S. Food and Drug Administration
GDP	Gross Domestic Product
HRQOL	Health-related Quality of Life
HGNC	HUGO Gene Nomenclature Committee
MeSH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
RDRD	Rare Disease Repurposing Database
UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

Part I

INTRODUCTION

The following chapter is an introduction to orphan diseases. The chapter starts out with an overview which includes general information regarding orphan diseases, such as the definitions and legislatives. Following the overview the socio-economic aspects (1.1) of orphan diseases, such as the cost of treatment and the revenue for pharma companies, will be briefly examined. After this the state of the *in silico* drug repurposing art in general and specific to orphan diseases (1.2) will be introduced. The last part of this chapter (1.3) will introduce validation strategies for *in silico* repurposing methods.

ORPHAN DISEASES

Orphan diseases are rare but numerous (up to 8,000 diseases have been described to date), affecting worldwide roughly 400 million people, of which 30 million are in Europe and 25 million in the US [1]. Specific definitions of orphan diseases are based on a low prevalence and vary slightly by country and organisation. An overview of relevant drug legislations and definitions for the prevalence of the disease, number of patients etc. can be found in Table 1, which originates from the paper ‘Orphan drugs: the regulatory environment’ by Franco [2].

In the EU the term *orphan disease* refers to conditions that ‘occur so infrequently that the cost of developing and bringing to the market a medicinal product to diagnose, prevent or treat the condition would not be recovered by the expected sales of the medicinal product; the pharmaceutical industry would be unwilling to develop the medicinal product under normal market conditions’, as stated in *REGULATION (EC) No 141/2000 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 1999 on orphan medicinal products* [3]. The prevalence is set at no more than 5 in 10,000 citizens. Some of these diseases might sound familiar, for example Cystic fibrosis or Huntington’s disease, the majority however are generally not well known, often having a patient population of fewer than a hundred. An orphan disease should by definition be very rare (therefore the terms *orphan* and *rare disease* are used synonymously), nevertheless medicinal products for life-threatening, seriously debilitating or serious and chronic conditions should be eligible for an orphan designation even when the prevalence is higher than indicated for a rare disease [3].

A considerable number of rare diseases has gene mutations as a cause, such as the deletion of a protein sequence or a mutation leading to a loss of function of a protein (e.g. an enzyme). Patients with such gene mutations are usually children suffering from a hereditary disease (such as Fragile X Syndrome). Hereditary diseases aside, examples for rare diseases are also rare types of cancer, or infections widely prevalent in some areas of the world but rare in others (such as Malaria) [4]. The term *neglected diseases* is at times used for the above mentioned infections, because in general they are highly prevalent in countries which do not have the means to pay for the treatment, meaning that they do not embody a profitable market for pharma companies and subsequently, as a research area, have been neglected throughout the years.

The incentives for the development of orphan medicinal products are numerous and range from protocol assistance, the partial waiving of fees to be paid to the European Medicines Agency and market exclusivity for a period of 10 years to support of research and development [3]. A detailed and regularly updated inventory of the existing incentives provided by the European Commission can be accessed at https://ec.europa.eu/health/sites/health/files/files/orphanmp/doc/orphan_inv_cwd_20160126.pdf [5]. A comparison of incentives provided by international legislations is shown in Table 3 in section 1.1.

Prevalence

The percentage of a population that is affected with a particular disease at a given time.

Definition

Orphan disease

An orphan disease affects less than 5 in 10,000 citizens. [EU]

An orphan disease affects less than 200,000 patients. [USA]

TABLE 1: Regulatory environments of orphan diseases [with permission from Elsevier (license number 4411470295517)]

There are several definitions of rare disease that are accepted by different countries according to their legislation or policy. The number of patients, necessary for being accepted as an orphan disease is in general related to the population size of these countries (with the exception of Singapore).

Country, region or organisation	Relevant orphan drug legislation	Definition	Population per country or region	Low prevalence of the disease	Number of patients necessary for being accepted as an orphan disease	Refs
WHO	–	A disease or a condition affecting 0.65–1 in 1000 inhabitants	–	6.5–10/10,000	N/A	
EU	Regulation EC No. 141/2000	A life-threatening or chronically debilitating condition affecting no more than 5 in 10,000 persons in the community; or life-threatening, seriously debilitating or serious chronic condition in EU and without incentives the Sponsor would develop the medicine; and there are no other satisfactory method of diagnostic, prevention or treatment of the condition	502,500,000	5/10,000	Fewer than 251,250	Demographics of the European Union Eurostat [6]
US	Orphan Drug Act of January 1983 and amendments	A disease or a condition, which affects fewer than 200,000 patients in US	311,864,524	6.4/10,000	Fewer than 200,000 or more 200,000 patients if the cost of development will not be recovered	US Census Bureau [7]
Japan	Law 145 – 10 August 1960 (revised in 1993)	A disease that affects fewer than 50,000 patients in Japan	127,950,000	3.9/10,000	Fewer than 50,000	Portal site of official Statistics of Japan [8]

TABLE 1 Continued: Regulatory environments of orphan diseases [with permission from Elsevier (license number 4411470295517)]

Country, region or organisation	Relevant orphan drug legislation	Definition	Population per country or region	Low prevalence of the disease	Number of patients necessary for being accepted as an orphan disease	Refs
Australia	Therapeutic Goods Act in 1989 (revised in 1997)	A disease that affects fewer than 2000 patients per year	22,663,156	$\approx 1/10,000$	Fewer than 2000	Australian Bureau of Statistics [9]
Canada	No specific legislation	Canada accepts the WHO definition	34,531,000	$\approx 1/10,000$	Fewer than 3300	Statistics Canada [10]
Singapore	Medicines Act Chapter 176, section 9)	A life-threatening and severely debilitating illness affecting fewer than 20,000 patients	5,100,000	$\approx 39.4/10,000$	Fewer than 20,000	National Population and Talent Division [11]
Taiwan	Rare disease and orphan drug act	Diseases with prevalence lower than formulated and publicity announced by the central competent authority, and recognised by the orphan drug committee	23,188,078	No specific number of patients is mentioned in the legislation	No specific number of patients is mentioned in the legislation	National Statistics Republic of China (Taiwan) [12], Rare Disease and Orphan Drug Act (Twain) [13]

1.1 SOCIO-ECONOMIC ASPECTS

The large number of orphan disease patients worldwide makes orphan diseases a key health priority, which is why there is a number of incentives provided to pharma companies before, during and after the development of orphan drugs, such as regulatory assistance, market exclusivity or a reduction of fees. A detailed comparison of these incentives is shown in Table 3. Indeed, before these incentives came into play, in the US for example fewer than ten drugs were approved by the FDA for orphan diseases [14], whereas currently the number amounts to 496 drugs with at least one orphan disease designation [15].

Overall the implementation of new orphan drug legislations led to a substantial increase in orphan designations as well as marketing approvals. The current number of orphan designations is 3,345 in the US and 1,146 in the EU. Marketing approval was given to 496 products in the US (as mentioned above) and 87 in the EU [16]. It is also worth mentioning that the regulatory success rate for orphan drugs that enter clinical testing equals 22%, which is twice as high as the rate for non-orphan drugs [14].

To date the number of drugs with orphan designation approved by the U.S. Food and Drug Administration (FDA) was highest in 2015 at 21, and was in line with a steady increase from 17 in 2014 and 9 in 2013 [17]. In 2016 the FDA approved 9 such products [18]. In 2015 over half of these drugs were cancer drugs, which reflects a trend to split cancer indications into small orphan sections and thus benefit from the incentives provided by the orphan drug legislation [4, 17]. Additionally, within the pharmaceutical industry, not only is there a trend to split indications, there is also a general trend to move towards rare disease drugs [18], due to the fact that these drugs can be quite profitable, commanding a market worth of nearly US\$90 billion annually [14].

A mid-level sale orphan drug such as Celgene's Vidaza[®] (a treatment for myelodysplastic syndrome) generates a revenue between \$100 and \$500 million per year [19]. However, blockbusters can reach revenues above \$500 million, as evidenced by Amgen's Sensipar[®] (a treatment for secondary hyperparathyroidism) which brought a revenue of \$808 million in 2011 [20]. Often when there are very few patients the cost of the treatment is especially high. The world's most expensive drug, Alexion Pharmaceutical's Soliris[®], is a treatment for paroxysmal nocturnal haemoglobinuria, which is a rare disease affecting approximately 5,000 individuals in the US [21]. It is priced at \$409,500 per patient per year [14, 22], posing a challenge for health insurance carriers.

Overall the estimation of the average value of an orphan drug's revenue over its competition-free lifespan is US\$1.36 billion, with a final average valuation of US\$818 million (revenue minus marketing costs etc.) [14].

The high cost of these treatments poses a financial and ethical problem. In the US a treatment is considered cost-effective if the cost of an additional year of life equals the annual Gross Domestic Product (GDP) per capita (\$40,000), and cost-ineffective when it equals three times the annual GDP (\$120,000). There is however no guideline for rare diseases on how to proceed when the treatment does exceed the annual GDP, given that often these treatments are the only ones

available to provide relief to patients. An example mentioned by Borski in ‘Economic aspects of rare diseases.’ is the treatment cost for a patient suffering from mucopolysaccharidosis type II, which amounted to 3.39 million PLN (€850,000) per year [23]. To cover the cost of this treatment the funds from health insurance premiums of approximately 1,066 people were needed ¹.

To study the economic aspects of orphan diseases in further detail, and also to assess the loss of the Health-related Quality of Life (HRQOL) of patients suffering from orphan diseases, the Social Economic Burden and Health-Related Quality of Life in Patients with Rare Diseases in Europe (BURQOL-RD) project began in April 2010, a 3-year project under the framework of the Second Programme of Community Action in the Field of Public Health [24, 25]. Its aim was to develop a model capable of quantifying the socio-economic burden and HRQOL of rare disease patients in Europe based on evidence on direct and indirect costs of ten rare diseases, which are shown in Table 2.

Cystic Fibrosis	Duchenne Muscular Dystrophy
Fragile X Syndrome	Haemophilia
Juvenile Idiopathic Arthritis	Mucopolysaccharidosis
Scleroderma	Prader-Willi Syndrome
Histiocytosis	Epidermolysis Bullosa

TABLE 2: BURQOL-RD’s ten rare diseases

The BURQOL-RD was a cost-of-illness study. This type of study measures the socio-economic burden of a disease and can potentially be used for public policy development when it comes to decisions regarding prioritisation and justification of health care as well as prevention policies. Estimations in cost-utility and cost-effectiveness can be made as well and are also helpful for policy makers. An effort was made to include rare and ultra-rare diseases, and diseases with and without existing pharmacological treatments.

The study comes to the conclusion that social costs (such as formal or informal care and loss of labour productivity) incurred by rare diseases are even higher than direct health care costs. The annual costs per patient vary considerably between the above mentioned diseases, with the highest mean yearly total cost per patient being incurred by Mucopolysaccharidosis, where the cost ranges from €24,520 in Hungary to €209,420 in Germany [25].

As evidenced by the low number of approved treatments and the large number of orphan diseases, the rare disease community can definitely benefit from new therapies that are safe and ideally also affordable. The workflow presented in this thesis is a step towards this goal.

¹ Derived from numbers which were published in ‘Economic aspects of rare diseases.’ by Borski [23].

TABLE 3: Incentives for the development of orphan drugs [with permission from Elsevier (license number 4411470295517)]

Comparison of incentives described in the orphan drug legislation or policies adopted by different regions						
	<i>Incentives to promote the development of orphan drugs in different countries or regions</i>					
	US-FDA	EU-EMA	Japan	Australia	Taiwan (Rare Disease and Orphan Drug Act)	Canada
Orphan drug legislation	Yes	Yes	Yes	Yes	Yes	No (policy)
Market exclusivity of an orphan drug	7 years	10 years (might be reduced to 6 years), plus 2 additional years for medicines that comply with the agreed paediatric investigate plan	10 years - extension of re-examination period	N/A (under discussion)	10 years + 2 years	N/A
How to break the market exclusivity for similar medicinal product	Clinical superiority and no <i>sameness</i>	Clinical superiority and no <i>sameness</i>	N/A	N/A	Clinical superiority and not <i>the same kind</i>	N/A
	Or the first MAH ^a is unable to provide sufficient quantities	Or the first MAH ^a is unable to provide sufficient quantities			Or the first MAH ^a is unable to provide sufficient quantities	
	Or consent of the first MAH ^a	Or consent of the first MAH ^a			Or consent of the first MAH ^a Or if the price of the first orphan is unreasonable	
Marketing authorisation review process	Allows sponsor to apply for accelerated review	Allows sponsor to apply for accelerated review	Fast track review	Allows sponsor to apply for accelerated review	N/A	Fast track review can be considered

TABLE 3 Continued: Incentives for the development of orphan drugs [with permission from Elsevier (license number 4411470295517)]

<i>Incentives to promote the development of orphan drugs in different countries or regions</i>						
	US-FDA	EU-EMA	Japan	Australia	Taiwan (Rare Disease and Orphan Drug Act)	Canada
Grants from regulatory competent authorities	Yes	Some incentives by member states and EC - Community research program - 7th framework program grant	Yes	No	Yes - determined by the central competent authority	Tax incentives
Fiancial incentives	50% federal tax credit for clinical research	Member states incentives	Tax Exemption Law 12% of expenses	No	Determined by the central competent authority	Tax incentives
Fee reduction for MAA^b	Yes	Fee reduction for an application 100% fee reduction for inspections	Tax reduction (16%)	Yes	Yes	Yes
Scientific advice (protocol assistance and/or consultation for development)	Yes	Yes - 100% fee reduction	Yes	No	Yes	Yes
Special incentives for SME (small and medium sized enterprise) sponsors	N/A	Waiver of market authorisation application fees and for post authorisation activities (50% reduction)	N/A	N/A	N/A	N/A
Regulatory assistance	Yes	Yes	Yes	Yes	Yes	Yes

^a Marketing Authorisation Holder

^b Marketing Authorisation Application

1.2 STATE OF THE DRUG REPURPOSING ART

Drug repurposing or repositioning are expressions which refer to the application of approved drugs to treat new indications. The usage of approved drugs for new diseases has several advantages. There is no need for a large infrastructure, because the research and development of the drug has already been done. Also, nearly half of the overall cost can in general be bypassed because no preclinical assessment is needed. Given that the starting point is usually an approved compound (known bioavailability and safety profile, proven manufacturing and formulation route, well characterized pharmacology), repurposing can additionally significantly reduce the risks associated with drug development (high failure rate of compounds, high cost, long timespan) as well as facilitate the entry of repurposed drugs into clinical phases more rapidly and at a lower cost [26, 27]. All of the above mentioned points are extremely advantageous when it comes to orphan diseases. However, most of the drugs that have been repositioned and are currently on the market are the result of serendipity, clearly showing a need for a more systematic approach for drug repositioning.

The art of orphan drug discovery in general has seen some progress in the last years, and the relationship between the pharmaceutical industry, academia and the patient community is currently very symbiotic. There are now quite a few resources devoted to orphan drug development, such as databases, disease-specific support groups, foundations, organisations, government programmes or private initiatives by pharma companies.

Patient advocacy alliances and organisations are for example the EURORDIS - Rare Diseases Europe², a non-profit alliance of 765 rare disease patient organisations from 69 countries, or NORD[®], the National Organisation for Rare Disorders³, a charity organisation with more than 260 patient organisation members. Their aim in general is to connect all stakeholders (patients, their families, physicians, researchers, governmental institutions, pharma companies etc.) and thus strengthen the patient voice and shape research, policies and patient services.

To foster collaboration internationally, there is the Rare Disease Research Consortium⁴, a coalition of 170 research sites around the world, whose goal is to produce treatments for rare diseases by 2020. On a national scale, the National Center for Advancing Translational Sciences⁵ has a specific programme called Therapeutics for Rare and Neglected Diseases⁶ to stimulate collaborations among NIH (National Institutes of Health)⁷ and academic scientists, nonprofit organisations and pharmaceutical and biotechnology companies working on rare and neglected diseases.

There are many academic centers involved in orphan disease drug discovery, research and development. For a partial list of academic centers in the US please

2 <https://www.eurordis.org/>, accessed on 06.10.2017

3 <https://rarediseases.org/>, accessed on 06.10.2017

4 www.irdirc.org, accessed on 06.10.2017

5 <https://ncats.nih.gov/index.php>, accessed on 06.10.2017

6 <https://ncats.nih.gov/trnd>, accessed on 06.10.2017

7 <https://www.nih.gov/>, accessed on 06.10.2017

consult Table 1 in ‘Orphan diseases: state of the drug discovery art’ by Volmar *et al.* [28]. Being that the work for this thesis has been done at the University of Vienna, I would also like to mention that Vienna itself has such a research center, the Vienna Center for Rare and Undiagnosed Diseases⁸.

When it comes to online resources for drug repurposing for orphan diseases there are quite a few, usually focusing on either the patients’ and physicians’ or the researchers’ needs, the latter of which are suitable for *in silico* drug repurposing approaches. During my search for suitable resources I came across a number of databases (briefly elaborated on below) which are of great value to the rare disease community, even though they are largely unsuitable for *in silico* drug repurposing. Databases for patients provide specific informations about rare diseases, contacts to physicians or health care centers as well as references to patient organisations pertaining to specific diseases. An example of a resource for patients is GARD, the Genetic and Rare Disease Information Center⁹. Resources of a more interactive nature, intended to help diagnose an orphan disease, are provided for physicians. Two such resources are specifically worth mentioning, one is FindZebra¹⁰, and the other is The Phenomizer¹¹. They are quite powerful search engines which are able to provide ideas for a diagnosis based on the entered symptoms or the phenotype. In The Phenomizer patient features can also be marked as mandatory or observed, thus narrowing down the diagnosis. For a thorough insight into computer based diagnostics of orphan diseases and corresponding platforms please consult ‘Computerunterstützte Diagnosefindung bei seltenen Erkrankungen’ by Müller *et al.* [31].

In the US the biggest online resource is the previously mentioned NORD[®], with a combination of information for patients, their families, clinicians, researchers, and the industry. In Europe there is Orphanet¹², a database with high-quality information on rare diseases for all stakeholders with the additional benefit of its own data repository called Orphadata¹³, which made this work possible in the first place because it provides datasets suitable for *in silico* drug repurposing. Orphadata provides for example datasets such as diseases and cross-references and diseases and their associated genes, both of which were utilised for this work. Further elaboration on Orphadata will follow in subsection 2.1.1.

When it comes to online resources for compounds that can be utilised for repurposing, there is currently a gap that urgently needs to be filled. Two such resources have been previously introduced in scientific publications, yet none of them is available any more. The first was the Pharmaceutical Assets Portal Project, whose aim was to collect compounds which were discontinued at various clinical stages and thus create a central repository available for academic scientists for drug repositioning [26, 34]. Such a resource would be of untold value to several drug discovery disciplines because with a proper annotation of why a drug was discontinued, drugs could be repurposed based on either the mech-

*I look for zebras because
other doctors rule out
all the horses.
– Dr. House*

⁸ <https://cerud.meduniwien.ac.at/>, accessed on 06.10.2017

⁹ <https://rarediseases.info.nih.gov/>, accessed on 06.10.2017

¹⁰ <http://www.findzebra.com/>, accessed on 06.10.2017 [29]

¹¹ <http://compbio.charite.de/phenomizer/> [30], accessed on 06.10.2017

¹² <http://www.orpha.net/> [32]

¹³ <http://www.orphadata.org/> [33]

anism of action or the side effects. In addition toxicological readouts could be performed. The second resource was a database made specifically for drug repurposing, the Rare Disease Repurposing Database¹⁴, which was introduced by the FDA [26]. The Rare Disease Repurposing Database included a list of 236 products that have received an orphan status designation, meaning that they were found to be potentially beneficial in treating a rare disease, even though they were not yet approved for that specific indication [35]. This database constituted the basis of this work's sensitivity-based validation effort. Additional information on the database can be found in the section Databases (2.1).

Quite a famous example for orphan drug repurposing is thalidomide (initially sold under the tradename Contergan), originally introduced as a sedative in the late 1950s. It was later prescribed for morning sickness and then subsequently withdrawn in 1961 because of its embryotoxicity (it led to malformations of limbs). In a classic case of serendipity in 1964 physician Jacob Sheskin at the University Hospital of Marseilles (France) administered thalidomide as a last resort to treat weeks of sleeplessness of a critically ill erythema nodosum leprosum (a complication of leprosy) patient suffering from severe pain. The day after the administration Sheskin observed that the patient's sores were healed and the pain was eliminated. Thalidomide was therefore revived in 1998 for the treatment of erythema nodosum leprosum and is currently also in use for a variety of other diseases [27]. The repurposing of thalidomide is a classic case of serendipity which led to one of the few drug repurposing success stories for orphan diseases. A handful of other examples can be found in Table 4 which originates from the paper 'Orphan/rare drug discovery through drug repositioning' by Muthyala [26]. The low number of approved drugs for orphan diseases as well as the fact that ideas for drug repurposing arise in most cases due to serendipity leave room for the implementation of systematic methods.

Aside from serendipity, ideas for drug repositioning can be based on informed insights (e.g. side effects) or derived from systematic analytical platforms [27]. Such a systematic analysis is backed by either of the following core principles. One is the "promiscuous" nature of drugs, meaning that a single drug can interact with multiple targets (proteins are often called targets in the context of drugs acting on them) and/or multiple pathways. In this case the compound is known and the effort is made to find a new target. The second principle focuses on the targets relevant to a disease and the fact that these can also play roles in other pathways, phenotypes or biological processes. Another thought voiced by Sardana *et al.* in 'Drug repositioning for orphan diseases' is that even though the links between targets and diseases and the links between targets and compounds are established, the full associations between a disease, its targets and respective compounds are often not clear, in the worst case even unavailable, but in the best case only "hidden" in a database or in the literature [27]. This allows for a triangulation of the information and has been the method of choice for this work.

14 <http://www.fda.gov/ForIndustry/DevelopingProductsforRareDiseasesConditions/HowtoapplyforOrphanProductDesignation/ucm216147.htm>, accessed on 06.10.2017

Examples of orphan drugs developed from repurposed drugs [with permission from Elsevier (license number 4411470913015)]		
Drug	Condition	Orphan disease
Duloxetine	Antidepressant	Urinary continence
Difluoromethylornithine	Cancer	Sleeping sickness
Minocycline	Antibiotic	Ischemic stroke
Gentamycin	Antibiotic	Duchenne muscular dystrophy
Thalidomide	Morning sickness	Multiple myeloma, leprosy
Amphotericin B	Antifungal	Leishmaniasis
Tamoxifen	Breast cancer	Bipolar disorder
Closantel	Veterinary antihelmintic	Onchocerciasis
HIV protease inhibitor	HIV-AIDS	Cancer
Sulindac	Cyclooxygenase inhibitor	Cancer
Arsenic trioxide	Syphilis	Acute promyelocytic leukemia
Cozaar	Blood pressure medication	Marfan syndrome
Viagra	Erectile dysfunction	Pulmonary hypertension
Memantine	Alzheimer's	Down syndrome
Arbaclofen	Cerebral palsy	Fragile X
Ivermectin	Antiparasitic (river blindness)	Malaria
Rituxan	Lymphoma	Rheumatoid arthritis
Miltefosin	Breast cancer	Leishmania
Tretinoin	Acne	Promyelocytic leukemia
Hydroxy urea	Cancer	Sickle cell disease

TABLE 4: Drug repurposing examples for orphan diseases

Examples of systematic methods for drug repurposing include a knowledge-based approach, where data that has been accumulated by the drug industry and academia is integrated and mined for non-explicit connections; a high-throughput screening of the pharmacopoeia against diverse targets to identify side-effects or possible therapeutic benefits; a screening with a disease as end-point, where drugs are screened against a phenotype of interest to discover on- and off-target effects as well as *in silico* methods based on pharmacophores and molecular similarity, networks and systems biology or machine learning [27, 36].

More detailed information on general as well as specific repurposing methods for orphan drug development can be found in 'Orphan diseases: state of the drug discovery art' by Volmar *et al.* [28], 'Drug repositioning for orphan diseases' by Sardana *et al.* [27] and 'In silico repositioning of approved drugs for rare and neglected diseases' by Ekins *et al.* [36].

1.3 VALIDATION STRATEGIES FOR COMPUTATIONAL DRUG REPURPOSING

A validation is a means to objectively determine the performance of a method. In the case of computational methods this can be quite difficult, especially if experimental data for the predictions cannot be obtained. Therefore for computational approaches such as the workflow presented here the validation is done

by comparing the computational results to existing biomedical knowledge. This comparison can for example be done with a literature search or comparison with databases such as ClinicalTrials.gov¹⁵. Within such a validation a good method would be able to consistently identify known associations between diseases and drugs. There are however no consistent best practices when it comes to a comparison of drug repurposing studies and the validation of drug repurposing methods, as stated in ‘A review of validation strategies for computational drug repositioning’ by Brown & Patel [37]. The authors identified three major types of validation strategies employed in studies in the computational drug repurposing field:

- Validation with a single example or case study of a single disease area (CSV)
- Sensitivity-based validation only (SV)
- Both sensitivity- and specificity-based validation (SSV)

The authors postulate that SV methods are superior to CSV methods because they assess the overall ability of the repositioning method instead of focusing on a few selected predictions in depth. In contrast an SSV approach requires information about falsely predicted drug indications (false positives). Often unannotated drug-indication pairs are marked as false positives which can be problematic. The choice of the annotation database substantially impacts the sensitivity and specificity estimates and marking the unannotated pairs as false positives suggests that all repositioning hypotheses generated by the respective method are false. This is of course counterintuitive when trying to predict novel associations.

Brown & Patel conclude that with the currently available data SV is the optimal choice for an analytic validation of repurposing studies.

15 <https://clinicaltrials.gov/>, accessed on 19.10.2017

Part II

MAIN BODY

The aim of this work was to create a workflow capable of proposing drug repurposing candidates for orphan diseases. This chapter provides detailed information about the methods that were used to create the workflow.

The first part is an introduction into the specific databases and platforms used for data retrieval within the workflow (2.1). The platform used for the creation of the workflow, the respective file formats and their means of retrieval as well as the implementation of text mining in KNIME[®] are described in section 2.2. The text mining was an additional data mining effort to streamline the process of assigning categories to disease-target associations, which were needed for a correct assignment of potentially interesting drug classes. The last part (2.3) describes the final configuration of the workflow in detail.

METHODS

The vision behind the workflow was to connect orphan diseases to approved compounds that might be repurposed for the given disease. Therefore the workflow needed to include a list of the aforementioned orphan diseases, a connecting step, and a list of potential candidates for repurposing. The connection was achieved through targets and the integration of databases which have information on disease-target and target-drug relationships. This approach is called triangulation because of the three cornerstones that are implemented and because of their interconnection which can also be seen as a triangle (depicted below in Figure 1).



FIGURE 1: Triangulation between diseases, targets and drugs

2.1 DATABASES

This section lists databases which were used for the triangulation in the order of their use within the workflow.

2.1.1 *Orphanet*

orphanet

Orphanet was chosen as the starting point because it is a resource dedicated to information about orphan diseases. It is an online platform which was established “in order to gather scarce knowledge on rare diseases so as to improve the diagnosis, care and treatment of patients with rare diseases” [32]. It provides high-quality information about roughly 9,000 rare diseases, including information about cross-references to other databases and information about disease-target relationships, both of which were used in this work. Orphanet can be accessed at <http://www.orpha.net/> and queried for information about for example diseases, orphan drugs, patient organisations, expert centres or ongoing research projects.

The beginning of an Orphanet entry for a disease, in this case for cystic fibrosis, is shown in Figure 2. Apart from the sections that are visible here, sections on the etiology of the disease, on diagnostic methods or on the management and treatment of the disease may be present.

Cystic fibrosis [Suggest an update](#)

Disease definition

Cystic fibrosis (CF) is a genetic disorder characterized by the production of sweat with a high salt content and mucus secretions with an abnormal viscosity.

ORPHA:586

Synonym(s):	Inheritance: Autosomal recessive	UMLS: C0010674
CF	Age of onset: All ages	MeSH: D003550
Mucoviscidosis	ICD-10: E84.0 E84.1 E84.8 E84.9	GARD: 6233
Prevalence: 1-9/100 000	OMIM: 219700	MedDRA: 10011762

Summary

Epidemiology

It is the most common genetic disorder among Caucasian children. The incidence varies between populations: the condition is considerably less common in Asian and African populations than in the white populations of Europe and North America, with variation within each country. The exact prevalence in Europe is unknown, but estimates range between 1/8,000 and 1/10,000 individuals.

Clinical description

FIGURE 2: Example of an Orphanet entry (screenshot)

All of the information present in such an entry is available for download from a dedicated site called Orphadata, accessible at <http://www.orphadata.org/>¹. Orphadata offers the download of thematically specialised files in XML² and some also in the JSON³ format. Some of these are freely accessible and for some of them Material Transfer Agreements have to be signed. For this work two of the freely accessible datasets were downloaded in the XML format. The first was the ‘Rare diseases and cross-referencing’ dataset⁴ and the second was the ‘Rare diseases with their associated genes’ dataset⁵. During the development and subsequent adjustment of the workflow for poster presentations and updates in the KNIME[®] Analytics Platform these files were updated regularly to reflect the newest available information on Orphanet. The versions corresponding to the final workflow and results described in this thesis are version 1.2.7 / 4.1.6 [2017-03-09] (*orientdb version*) for the ‘Rare diseases with their associated genes’ dataset, and version 1.2.7 / 4.1.6 [2017-03-09] (*orientdb version*) for the ‘Rare diseases and cross-referencing’ dataset.

1 Orphadata. Orphadata: Free access data from Orphanet. ©, INSERM 1997 Accessed on 19.01.2018. <<http://www.orphadata.org/cgi-bin/index.php/>>.

2 <https://en.wikipedia.org/wiki/XML>, accessed on 19.01.2018

3 <https://www.json.org/>, accessed on 19.01.2018

4 Orphadata. Rare diseases and cross-referencing Accessed on 31.10.2018. <http://www.orphadata.org/data/xml/en_product1.xml>.

5 Orphadata. Rare diseases with their associated genes Accessed on 31.10.2018. <http://www.orphadata.org/data/xml/en_product6.xml>.

The ‘Rare diseases and cross-referencing’ dataset was used as the starting point in the workflow because it provides cross-references to other database identifiers for the diseases listed in Orphanet. Every disease listed in Orphanet has a unique Orphanet identifier, this identifier however could not be used in the workflow because it is not included in the Open PHACTS Discovery Platform, which was used to execute the second step of the triangulation. The cross-referencing however provided an ideal solution because Unified Medical Language System (UMLS)⁶ identifiers are cross-referenced in the dataset and they are also included in the Open PHACTS Discovery Platform, meaning that based on these identifiers, links to targets can be retrieved.

2.1.2 Open PHACTS Discovery Platform



The Open PHACTS Discovery Platform⁷ is a data integration platform which was built in collaboration with a large consortium of major academic and commercial organisations within the frame of an Innovative Medicines Initiative⁸ project. Currently the Open PHACTS Discovery Platform is maintained by the Open PHACTS Foundation⁹.

At its core, the Open PHACTS (Open Pharmacological Concept Triple Store) Discovery Platform is a triplestore¹⁰ of semantically linked databases. Open PHACTS assimilates data - in the form subject-predicate-object - provided by publicly-available databases such as DisGeNET (see also 2.1.3), DrugBank (see also 2.1.4) or ChEMBL in the RDF¹¹ format [43, 44]. The information stored in triples might look like this: *Acetylsalicylic acid inhibits Prostaglandin G/H synthase 1*. The subject here is *Acetylsalicylic acid*, the predicate is *inhibits* and the object is *Prostaglandin G/H synthase 1*. In this manner all sorts of associations can be stored. An example visualisation of the DisGeNET RDF is shown in Figure 3 [43].

Open PHACTS provides an Application Programming Interface (API) which is suitable for automated data retrieval within workflow management systems. API calls are described in detail in subsection 2.2.2. The Open PHACTS API version 2.2 was used in the final workflow.

The databases within Open PHACTS which were relevant to this work are DisGeNET (2.1.3) and DrugBank (2.1.4). The disease-target links for the triangulation originate from DisGeNET, and the target-drug links originate from DrugBank.

6 NIH. *Unified Medical Language System* Accessed on 19.01.2018. <<https://www.nlm.nih.gov/research/umls/>>.

7 Open PHACTS Foundation. *Open PHACTS Discovery Platform* Accessed on 19.10.2017. <<https://www.openphacts.org/>>.

8 <http://www.imi.europa.eu/projects-results/project-factsheets/open-phacts>, accessed on 19.01.2018

9 Open PHACTS Foundation. *Registered Charity* Accessed on 19.10.2017. <<http://www.openphactsfoundation.org/>>.

10 <https://en.wikipedia.org/wiki/Triplestore>, accessed on 19.01.2018

11 Resource Description Framework, <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>, accessed on 19.01.2018

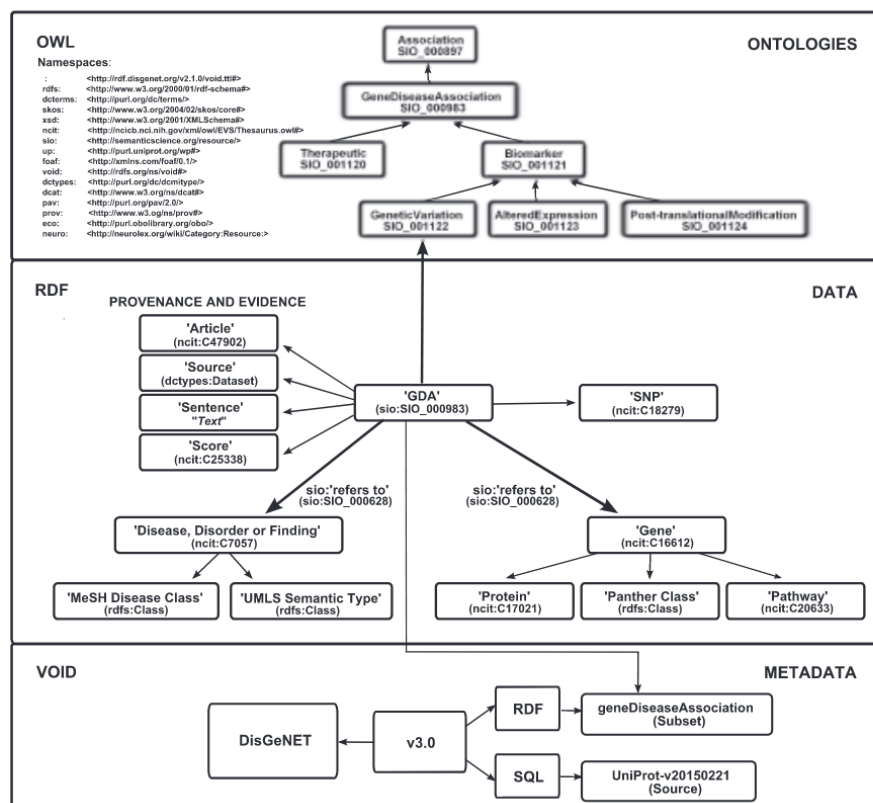


FIGURE 3: DisGeNET RDF [by permission of Oxford University Press (license number 4411430605272)]



2.1.3 DisGeNET

DisGeNET¹² is a platform with one of the largest collections of genes and variants associated with human diseases. DisGeNET integrates human gene-disease associations from several expert curated databases as well as text-mining derived associations [46, 47]. Scores as well as references are provided for all of the associations. Detailed information on the original data sources and the scoring function is provided on the website¹³.

In the workflow, data from DisGeNET was used to connect diseases to targets. Based on the UMLS identifiers lists of targets for the respective diseases were retrieved. The list of targets as seen in DisGeNET is shown in Figure 4.

The UMLS identifier is listed directly under the name of the disease, then there is additional information and several other identifiers, and at the bottom there is the beginning of the list of related genes (the list in the original entry continues further than shown in Figure 4). The cystic fibrosis transmembrane conductance regulator is the highest ranking gene here with a large number of references for the association (1,596 PubMed identifiers). The PubMed identifiers are used in the text mining approach for the validation of results, further described in subsection 2.2.3.

¹² Integrative Biomedical Informatics Group. DisGeNET Accessed on 23.10.2017. <http://www.disgenet.org/>.

¹³ <http://www.disgenet.org/web/DisGeNET/menu/dbinfo>, accessed on 23.10.2017

Cystic Fibrosis

UMLS CUI C0010674

Type disease

MeSH Class Congenital, Hereditary, and Neonatal Diseases and Abnormalities ; Digestive System Diseases ; Respiratory Tract Diseases

MeSH D003550

OMIM 219700

Semantic Type Disease or Syndrome

Phenotypes null

Disease Ontology genetic disease

Top 10 gene associations for this disease

Top scored gene associations for this disease.

Gene	Gene Name	Score	PMIDs
CFTR	cystic fibrosis transmembrane conductance regulat... >	1	1596
TGFB1	transforming growth factor beta 1	0.433	17
DCTN4	dynactin subunit 4	0.401	2
SCNN1B	sodium channel epithelial 1 beta subunit	0.293	13
SCNN1G	sodium channel epithelial 1 gamma subunit	0.212	6
SCNN1A	sodium channel epithelial 1 alpha subunit	0.207	9
TNFRSF1A	TNF receptor superfamily member 1A	0.205	1
CLCA4	chloride channel accessory 4	0.204	4
STX1A	syntaxin 1A	0.2	1
CXCL8	C-X-C motif chemokine ligand 8	0.037	65

FIGURE 4: Example of a DisGeNET entry (screenshot)

2.1.4 DrugBank



DrugBank¹⁴, a database with comprehensively annotated information about drugs, is suitable for the mining of drug data or for combining drug data across different databases. Compounds in the database are extensively annotated with information about their nomenclature, chemistry, structure, stages of approval (e.g. investigational, approved), function, action, pharmacology and pharmacokinetics [49]. The DrugBank entry for the target mentioned in the section on DisGeNET (2.1.3) - the cystic fibrosis transmembrane conductance regulator - is shown in Figure 5.

¹⁴ Wishart Research Group. *DrugBank* Accessed on 23.10.2017. <<https://www.drugbank.ca/>>.

DRUGBANK Browse Search Downloads About Help Contact Us

Cystic fibrosis transmembrane conductance regulator

DETAILS

Name Cystic fibrosis transmembrane conductance regulator

Kind protein

Organism Human

Polypeptides

NAME	UNIPROT ID
Cystic fibrosis transmembrane conductance regulator	P13569 Details

DRUG RELATIONS

Drug Relations Show 10 entries Search

DRUGBANK ID	NAME	DRUG GROUP	PHARMACOLOGICAL ACTION?	ACTIONS	DETAILS
DB00171	ATP	investigational, nutraceutical	unknown		Details
DB00887	Bumetanide	approved	unknown		Details
DB02587	Colforsin	experimental, investigational	unknown		Details
DB04941	Crofelemer	approved	yes		Details
DB04522	Dextrofosferine	experimental	unknown		Details
DB09213	Dexibuprofen	approved, investigational	unknown		Details

FIGURE 5: Example of a DrugBank entry (screenshot)

Under ‘DRUG RELATIONS’ a list of drugs which act on the protein is provided. The DrugBank ID of the drug is provided, its name, the stage of approval and the action of the drug. The action is necessary for the final validation step of the results from the triangulation.

The DrugBank data can be accessed from the online interface [50] in the download section (XML, free for non-commercial use, registration is required) and via an API (subject to a charge). An example of an XML entry of a drug is shown in Listing 3 in section 2.2.1. For the final workflow and the corresponding results described in chapter 3 the DrugBank XML version 5.0.11 (released on 20.12.2017) was used.

2.1.5 Rare Disease Repurposing Database

In 2011 the FDA published a paper on their Rare Disease Repurposing Database (RDRD) of approved drugs with the potential to be repurposed for the treatment of orphan diseases [35]. The database was meant to facilitate drug repurposing efforts and to speed up the delivery of new therapies. The database listed 236 products which have received orphan status designation and were found to be potentially beneficial for an orphan disease even though they did not yet have a marketing approval for said disease. At the beginning of this work the database was still accessible, which is why the information in the database was used for the validation effort of this approach (as described in section 1.3). During 2017 the database was however discontinued.

The data was available in the form of three Excel sheets, one for products with an approval for a common indication, one for products with an approval for a rare indication, and one for products with marketing approvals for both

common and rare indications. In Table 5 and 6 excerpts from the RDRD tables are shown. The product name, indication, sponsors, trade names etc. are listed here. The validation effort in which the RDRD data was used is described in detail in chapter 3 in section 3.2.

TABLE 5: Excerpt from Table 1 in the Rare Disease Repurposing Database [by permission of Oxford University Press (license number 4411460726513)]

Table 1: Orphan-designated products with at least one marketing approval for a common disease indication						
Product name	Orphan status designation date	Designated but not yet approved orphan indications	Sponsors holding orphan designations	Trade names ^a	Approved common disease indications ^b	Sponsors of approved products ^a
Adenosine	1 August 1989	For use in conjunction with BCNU in the treatment of brain tumors	Medco Research, Inc	Multiple brands	Conversion to sinus rhythm of paroxysmal supraventricular tachycardia, including that associated with accessory bypass tracts (Wolff-Parkinson-White syndrome)	Multiple sponsors
Albuterol	12 March 2002	Prevention of paralysis due to spinal cord injury	MotoGen, Inc	Multiple brands	For the prevention and relief of bronchospasm in patients with reversible obstructive airway disease, and for the prevention of exercise-induced bronchospasm	Multiple sponsors
Alefacept	19 December 2007	For use as prophylaxis of rejection in patients receiving allogenic solid organ transplants	Astellas Pharma Global Development, Inc	Amevive	Treatment of adult patients with moderate to severe chronic plaque psoriasis who are candidates for systemic therapy or phototherapy	Astellas Pharma US, Inc

^a If there are more than two trade names and/or sponsors for the same approved indication, 'multiple brands' or 'multiple sponsors' are used instead of listing all the brands/sponsors.

^b Any information about the product after the approval for marketing is not included in this database.

TABLE 6: Excerpt from Table 2 in the Rare Disease Repurposing Database [by permission of Oxford University Press (license number 4411460726513)]

Table 2: Orphan-designated products with at least one marketing approval for a rare disease indication						
Product name	Orphan status designation date	Designated but not yet approved orphan indications	Sponsors holding orphan designations	Trade names	Approved common disease indications	Sponsors of approved products ^a
111 Indium pentetreotide	10 June 1999	Treatment of somatostatin receptor positive neuroendocrine tumors	Louisiana State University Medical Center Foundation	Octreoscan	Indium In-111 pentetreotide is an agent for the scintigraphic localization of primary and metastatic neuroendocrine tumors bearing somatostatin receptors	Mallinckrodt
	16 June 2006	Treatment of neuroendocrine tumors	Radioisotope Therapy of America			
Alglucerase injection	21 July 1995	Replacement therapy in patients with Typer II and III Gaucher's disease	Genzyme Corporation	Ceredase	Use as a long-term enzyme replacement therapy for children, adolescents and adult patients with a confirmed diagnosis of Type I Gaucher disease who exhibit signs and symptoms that are severe enough to result in one or more of the following conditions: (i) moderate to severe anemia; (ii) thrombocytopenia with bleeding tendency; (iii) bone disease; (iv) significant hepatomegaly or splenomegaly	Genzyme Corporation

2.1.6 *repoDB*

In 2017 Brown & Patel introduced repoDB, according to them a gold standard database, with the aim of unifying method validation in the *in silico* drug repurposing field [51]. The database consists of both true positives (approved disease-drug combinations), and true negatives (failed disease-drug combinations), meaning that it can be used to validate connections between diseases and drugs proposed by an *in silico* drug repurposing method.

Approved indications in the repoDB originate from DrugCentral¹⁵ and failed indications originate from the Aggregate Analysis of ClinicalTrials.gov Database¹⁶. The repoDB is publicly available and can be accessed at <http://apps.chiragjpgroup.org/repoDB/>. The full repoDB dataset can be downloaded from the platform and example entries from the dataset are shown in Table 7. The dataset provides the name of the drug, its DrugBank identifier, the name of the indication, its UMLS identifier, the ClinicalTrials.gov identifier of the clinical study the entry is based on, the status of the drug, the phase of the study, and in some cases additional details regarding the study.

The database was used to cross-check with the results from the workflow in an effort to validate the associations proposed by the workflow. The results of this validation effort are discussed in detail in chapter 3.

¹⁵ <http://drugcentral.org/>, accessed on 21.01.2018

¹⁶ <https://aact.ctti-clinicaltrials.org/>, accessed on 21.01.2018

TABLE 7: Example entries in the repoDB dataset

drug_name	drug_id	ind_name	ind_id	NCT	status	phase	DetailedStatus
Potassium Citrate	DB09125	Calcium renal calculus	C1959799	NA	Approved	NA	NA
Dextroamphetamine	DB01576	Cocaine-Related Disorders	C0236736	NCT00000305	Terminated	Phase 1	Original P.I. left the institution
Cycloserine	DB00260	Schizophrenia	C0036341	NCT00000372	Withdrawn	Phase 3	Pairing D-Cycloserine with Clozapine was found to worsen negative side effects in patients with Schizophrenia, so the study was suspended.
Calcitriol	DB00136	Rickets	C0035579	NCT00001151	Terminated	Phase 2	Terminated because of lack of drug supply
Daclizumab	DB00111	Aplastic Anemia	C0002874	NCT00001962	Terminated	Phase 2	DSMB recommended closing the study and publishing the results
Cyclosporine	DB00091	leukemia	C0023418	NCT00003056	Terminated	Phase 3	Interim analysis indicated study should be terminated

2.1.7 *EMBL-EBI webservices*

The European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI) web services¹⁷ were used to retrieve publication abstracts (needed for the text mining approach which is described in subsection 2.2.3). EMBL¹⁸ is an intergovernmental organisation and the EBI is its '*home for big data in biology*'¹⁹.

The EMBL-EBI web services provide programmatic access to a large number of sources [55, 56], among others to Europe PubMed Central²⁰. The specifics of the retrieval of the abstracts are described in detail in section 2.3. Europe PMC²¹ is a repository which provides access to life science articles, books, patents as well as clinical guidelines. It includes resources from PubMed²² and PubMed Central²³. Where legally applicable Europe PMC provides abstracts as well as the full text of articles.

17 EMBL-EBI. *EMBL-EBI web services* Accessed on 26.04.2018. <<https://www.ebi.ac.uk/seqdb/confluence/display/WEBSERVICES/EMBL-EBI+Web+Services>>.

18 European Molecular Biology Laboratory Accessed on 26.04.2018. <<https://www.embl.org/>>.

19 European Bioinformatics Institute Accessed on 26.04.2018. <<https://www.ebi.ac.uk/>>.

20 Europe PMC. *Articles RESTful API - RESTful Web Service* Accessed on 26.04.2018. <<https://europepmc.org/RestfulWebService>>.

21 Europe PMC Funders' Group. *Europe PubMed Central* Accessed on 26.04.2018. <<https://europepmc.org/>>.

22 National Center of Biotechnology Information. *PubMed* Accessed on 26.04.2018. <<https://www.ncbi.nlm.nih.gov/pubmed>>.

23 National Center of Biotechnology Information. *PubMed Central*[®] Accessed on 26.04.2018. <<https://www.ncbi.nlm.nih.gov/pmc/>>.

2.2 KNIME[®] ANALYTICS PLATFORM

The KNIME[®] Analytics Platform is a freely available workflow management system. At the time of writing the final workflow was compatible with version 3.6.1 of the KNIME[®] Analytics Platform. Workflow management systems in general are modular environments for an interactive execution of a pipeline, meaning they are an infrastructure for the set-up, performance and monitoring of a defined sequence of tasks. There is a large number of available workflow management systems^{24, 25}. Some of these workflow management systems are freely available (such as the KNIME[®] Analytics Platform²⁶) and some are available commercially (such as Pipeline Pilot²⁷), each with its benefits and drawbacks [65].

The KNIME[®] Analytics Platform in specific is a data exploration platform with a graphical user interface which enables the user to visually create data flows, to execute some or all steps of an analysis and to investigate the results through interactive views [66]. It has a history of use for the mining of biomedical and chemical data and there are several community contributions addressing specific databases, accessing external software or performing dedicated cheminformatic tasks [67].

There are many tutorials^{28, 29} and example workflows³⁰ available which allow for an easy introduction to KNIME.

As KNIME[®] represents data internally in a specific format (the KNIME[®] table), as a first step the user always has to import their data using one of

FIGURE 6: Basic KNIME[®] workflow [63]

the reader nodes (the nodes in KNIME[®] are visualised units of data processing which can be interconnected). Afterwards the data can be transformed and mined with the use of nodes dedicated to data transformation. The last steps of a workflow are usually the visualisation and the storage or the deployment of the data. In Figure 6 a very basic workflow with the aforementioned steps is shown.

The workflows are created in the KNIME[®] workbench (depicted in Figure 7) which is sectioned into dedicated areas. By dragging and dropping desired nodes

24 Tomasso, P. D. *Awesome Pipeline* Accessed on 27.04.2018. <<https://github.com/pditommaso/awesome-pipeline>>.

25 Crusoe, M. R. *Computational Data Analysis Workflow Systems* Accessed on 27.04.2018. <<https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems>>.

26 KNIME AG. *KNIME[®] Analytics Platform* Accessed on 27.04.2018. <<https://www.knime.com/knime-analytics-platform>>.

27 BIOVIA. *Pipeline Pilot* Accessed on 27.04.2018. <<http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/>>.

28 KNIME AG. *Getting Started* Accessed on 27.04.2018. <<https://www.knime.com/knime>>.

29 KNIME GmbH. *KNIME Quickstart Guide* Accessed on 27.04.2018. <https://files.knime.com/sites/default/files/inline-images/KNIME_quickstart.pdf>.

30 KNIME AG. *Example Workflows* Accessed on 27.04.2018. <<https://www.knime.com/example-workflows>>.

into the *Workflow Editor* area and by interconnecting them the user can build a customised workflow.

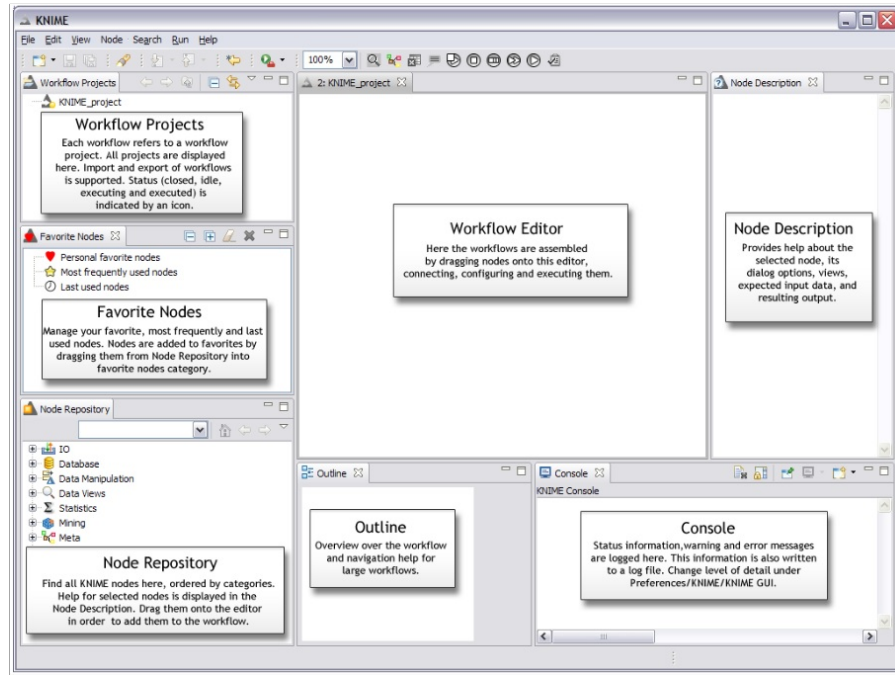


FIGURE 7: KNIME® workbench [69]

The user first has to install the core KNIME® implementation. For this work especially the *XML Reader* and the *XPath* nodes (shown in Figure 8) were essential for the retrieval and mining of data as described in subsection 2.2.1.

As mentioned before, KNIME® uses an internal data representation format, meaning that whichever type of data the user wants to analyse has to be imported into KNIME®. In my case the import consisted of two types of retrieval, a download of the entire database and access through API calls, both of which are described in detail in section 2.2.2. All of the data I worked with was in XML format which is described in detail in subsection 2.2.1.

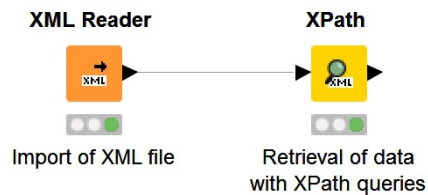


FIGURE 8: KNIME® XML Reader and XPath nodes (screenshot)

2.2.1 Data Format: XML

XML is a text format for storing data and can roughly be described as a tree in which entries relate to each other in a parent-child fashion with values as well as attributes. In some cases the XML includes a prolog which consists of an XML Declaration and a Document Type Declaration. The prolog is optional, but if one is included, it has to be positioned at the very beginning. An example of an XML file are shown in Listing 1:

A node (a data element) is introduced and closed in a certain syntax, as seen for example for `<heading>` which is introduced as `<heading>` and then closed as

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <note date="2017-11-22">
3   <from>Jana</from>
4   <to>Reader</to>
5   <body>
6     <heading>Chemist Joke</heading>
7     <line1>Two atoms are walking along.</line1>
8     <line2>One of them says:</line2>
9     <line3>"Oh no, I think I lost an electron."</line3>
10    <line4>"Are you sure?"</line4>
11    <line5>"Yes, I'm positive."</line5>
12  </body>
13 </note>

```

LISTING 1: Example XML

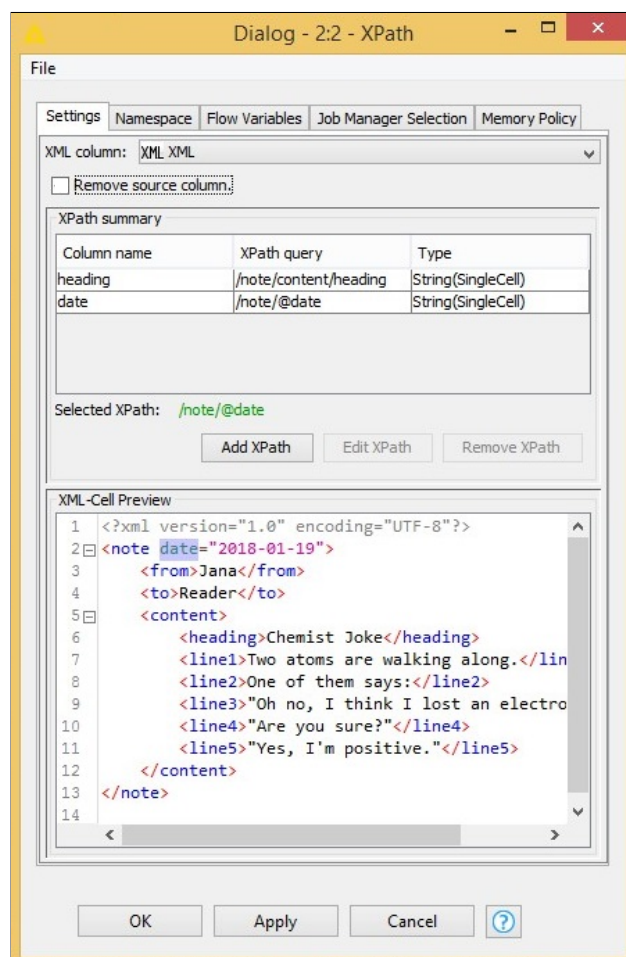
</heading>. The value of the node <heading> is 'Chemist Joke'. The content of a given node are all entries which are between the node's beginning and end. The only node with an attribute in this example is the node <note>, which has the attribute "date" with the corresponding value '2017-11-22'. As to the relationships, the first parent node in this case is <note>, with the children <from>, <to> and <body>, and the descendants <heading> and <line1> to <line5>. The second parent node is <body>, whose children are <heading> and <line1> to <line5>. Children are in turn siblings to each other. This syntax makes it possible to query an XML document for specific information.

The query language for XML is XPath³¹ and it is worth mentioning that even though there are different versions of XPath available, the KNIME[®] XPath node supports solely version 1.0. KNIME[®] provides an example workflow demonstrating the use of XPath nodes for the processing of XML files³².

An introduction to composing XPath queries can be found for example at https://www.w3schools.com/xml/xpath_intro.asp. An XPath query for the descendant <heading> in our example would be `/note/body/heading`, which takes the succession of nodes into account. Attributes on the other hand are addressed with an @ sign and the corresponding attribute name, so in our example a query for the date would be `/note/@date`. The configuration of the XPath node in KNIME[®] and the results for the given XPaths are shown in Figure 9.

31 <https://www.w3.org/TR/xpath/>, accessed on 27.04.2018

32 KNIME[®]. *Basic XML processing with XPath* Accessed on 27.04.2018. <<https://www.knime.com/nodeguide/data-access/structured-data/basic-xml-processing-with-xpath>>.



(A) XPath node configuration

Output Table - 2:2 - XPath

File Hilite Navigation View

Table "default" - Rows: 1 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	XML XML	S heading	S date
Row0	<?xml version="1.0" encoding="UTF-8"?> <note date="2018-01-19"> <from>Jana</from> <to>Reader</to> <content> <heading>Chemist Joke</heading> <line1>Two atoms are walking along.</line1> <line2>One of them says:</line2> <line3>"Oh no, I think I lost an electro <line4>"Are you sure?"</line4> <line5>"Yes, I'm positive."</line5> </content> </note>	Chemist Joke	2018-01-19

(B) XPath node results

FIGURE 9: XPath in KNIME® (screenshots)

Native data is of course often much more complex. As mentioned before, datasets from Orphanet (2.1.1) and the DrugBank database (2.1.4) were downloaded in the XML file format. These files were then imported into KNIME[®] with the XML Reader nodes. In Listing 2 an example of an Orphanet entry is shown. This is the original XML entry for the cross-referencing of cystic fibrosis (the Orphanet entry for cystic fibrosis is shown in Figure 2). The part of the XML entry which refers to the UMLS identifier is highlighted in yellow (lines 48-61). The UMLS identifier (present in line 50 directly below the `<Source>UMLS</Source>` entry) was retrieved with the XPath node using the following XPath: `/Disorder/ExternalReferenceList/ExternalReference/Source[.="UMLS"]/following-sibling::Reference`. With this syntax it was specified that the value for `<Source>` had to be 'UMLS' (there are many different cross-references and therefore `<Source>` nodes present in the XML, e.g. MedDRA identifiers) and that the retrieved value was to be the value of the following sibling called `<Reference>` instead of one of the many other `<Reference>` entries.

```

1  <?xml version="1.0" encoding="UTF-8"?>
2  <Disorder id="49">
3    <OrphaNumber>586</OrphaNumber>
4    <ExpertLink lang="en">http://www.orpha.net/consor/cgi-bin/
      ↪ OC_Exp.php?lng=en&Expert=586</ExpertLink>
5    <Name lang="en">Cystic fibrosis</Name>
6    <DisorderFlagList count="1">
7      <DisorderFlag id="475">
8        <Label>on-line</Label>
9      </DisorderFlag>
10   </DisorderFlagList>
11   <SynonymList count="2">
12     <Synonym lang="en">CF</Synonym>
13     <Synonym lang="en">Mucoviscidosis</Synonym>
14   </SynonymList>
15   <DisorderType id="21394">
16     <OrphaNumber>377788</OrphaNumber>
17     <Name lang="en">Disease</Name>
18   </DisorderType>
19   <ExternalReferenceList count="8">
20     <ExternalReference id="3746">
21       <Source>OMIM</Source>
22       <Reference>219700</Reference>
23       <DisorderMappingRelation id="21527">
24         <OrphaNumber>377807</OrphaNumber>
25         <Name lang="en">E (exact mapping (the terms and the concepts are
          ↪ equivalent))</Name>
26       </DisorderMappingRelation>
27       <DisorderMappingICDRelation>
28       </DisorderMappingICDRelation>
29       <DisorderMappingValidationStatus id="21611">
30         <OrphaNumber>377819</OrphaNumber>
31         <Name lang="en">Validated</Name>
32       </DisorderMappingValidationStatus>
33     </ExternalReference>
34     <ExternalReference id="104715">
35       <Source>MeSH</Source>
36       <Reference>D003550</Reference>
37       <DisorderMappingRelation id="21527">
38         <OrphaNumber>377807</OrphaNumber>
39         <Name lang="en">E (exact mapping (the terms and the concepts are
          ↪ equivalent))</Name>
40       </DisorderMappingRelation>
41       <DisorderMappingICDRelation>

```

```

42     </DisorderMappingICDRelation>
43     <DisorderMappingValidationStatus id="21611">
44         <OrphaNumber>377819</OrphaNumber>
45         <Name lang="en">Validated</Name>
46     </DisorderMappingValidationStatus>
47 </ExternalReference>
48 <ExternalReference id="104716">
49     <Source>UMLS</Source>
50     <Reference>C0010674</Reference>
51     <DisorderMappingRelation id="21527">
52         <OrphaNumber>377807</OrphaNumber>
53         <Name lang="en">E (exact mapping (the terms and the concepts are
           ↳ equivalent))</Name>
54     </DisorderMappingRelation>
55     <DisorderMappingICDRelation>
56     </DisorderMappingICDRelation>
57     <DisorderMappingValidationStatus id="21611">
58         <OrphaNumber>377819</OrphaNumber>
59         <Name lang="en">Validated</Name>
60     </DisorderMappingValidationStatus>
61 </ExternalReference>
62 <ExternalReference id="104717">
63     <Source>MedDRA</Source>
64     <Reference>10011762</Reference>
65     <DisorderMappingRelation id="21527">
66         <OrphaNumber>377807</OrphaNumber>
67         <Name lang="en">E (exact mapping (the terms and the concepts are
           ↳ equivalent))</Name>
68     </DisorderMappingRelation>
69     <DisorderMappingICDRelation>
70     </DisorderMappingICDRelation>
71     <DisorderMappingValidationStatus id="21611">
72         <OrphaNumber>377819</OrphaNumber>
73         <Name lang="en">Validated</Name>
74     </DisorderMappingValidationStatus>
75 </ExternalReference>
76 <ExternalReference id="104719">
77     <Source>ICD-10</Source>
78     <Reference>E84.0</Reference>
79     <DisorderMappingRelation id="21541">
80         <OrphaNumber>377809</OrphaNumber>
81         <Name lang="en">BTNT (broader term maps to a narrower term)</Name>
82     </DisorderMappingRelation>
83     <DisorderMappingICDRelation id="21583">
84         <OrphaNumber>377815</OrphaNumber>
85         <Name lang="en">Specific code (The term has its own code in the
           ↳ ICD10)</Name>
86     </DisorderMappingICDRelation>
87     <DisorderMappingValidationStatus id="21611">
88         <OrphaNumber>377819</OrphaNumber>
89         <Name lang="en">Validated</Name>
90     </DisorderMappingValidationStatus>
91 </ExternalReference>
92 <ExternalReference id="104720">
93     <Source>ICD-10</Source>
94     <Reference>E84.1</Reference>
95     <DisorderMappingRelation id="21541">
96         <OrphaNumber>377809</OrphaNumber>
97         <Name lang="en">BTNT (broader term maps to a narrower term)</Name>
98     </DisorderMappingRelation>
99     <DisorderMappingICDRelation id="21583">
100         <OrphaNumber>377815</OrphaNumber>
101         <Name lang="en">Specific code (The term has its own code in the
           ↳ ICD10)</Name>
102     </DisorderMappingICDRelation>

```

```

103     <DisorderMappingValidationStatus id="21611">
104         <OrphaNumber>377819</OrphaNumber>
105         <Name lang="en">Validated</Name>
106     </DisorderMappingValidationStatus>
107 </ExternalReference>
108 <ExternalReference id="104721">
109     <Source>ICD-10</Source>
110     <Reference>E84.8</Reference>
111     <DisorderMappingRelation id="21541">
112         <OrphaNumber>377809</OrphaNumber>
113         <Name lang="en">BTNT (broader term maps to a narrower term)</Name>
114     </DisorderMappingRelation>
115     <DisorderMappingICDRelation id="21583">
116         <OrphaNumber>377815</OrphaNumber>
117         <Name lang="en">Specific code (The term has its own code in the
            ↪ ICD10)</Name>
118     </DisorderMappingICDRelation>
119 </ExternalReference>
120 <ExternalReference id="104722">
121     <Source>ICD-10</Source>
122     <Reference>E84.9</Reference>
123     <DisorderMappingRelation id="21541">
124         <OrphaNumber>377809</OrphaNumber>
125         <Name lang="en">BTNT (broader term maps to a narrower term)</Name>
126     </DisorderMappingRelation>
127     <DisorderMappingICDRelation id="21583">
128         <OrphaNumber>377815</OrphaNumber>
129         <Name lang="en">Specific code (The term has its own code in the
            ↪ ICD10)</Name>
130     </DisorderMappingICDRelation>
131 </ExternalReference>
132 </ExternalReferenceList>
133 <DisorderDisorderAssociationList count="0">
134 </DisorderDisorderAssociationList>
135 <TextualInformationList count="6">
136     <TextualInformation id="1178" lang="en">
137         <TextSectionList count="1">
138             <TextSection id="100757" lang="en">
139                 <TextSectionType id="16907">
140                     <Name lang="en">Disease definition</Name>
141                 </TextSectionType>
142                 <Contents>Cystic fibrosis (CF) is a genetic disorder characterized by the
143                     ↪ production of sweat with a high salt content and mucus secretions
144                     ↪ with an abnormal viscosity.</Contents>
145             </TextSection>
146         </TextSectionList>
147     </TextualInformation>
148     <TextualInformation id="85778" lang="en">
149         <TextSectionList count="0">
150 </TextSectionList>
151 </TextualInformation>
152     <TextualInformation id="85429" lang="en">
153         <TextSectionList count="0">
154 </TextSectionList>
155 </TextualInformation>
156     <TextualInformation id="85428" lang="en">
157         <TextSectionList count="0">
158 </TextSectionList>
159 </TextualInformation>
160 </TextualInformationList>
161 </TextualInformationList>
162 </TextualInformationList>
163 </TextualInformationList>

```

```

164     </TextSectionList>
165   </TextualInformation>
166   <TextualInformation id="67855" lang="en">
167     <TextSectionList count="0">
168     </TextSectionList>
169   </TextualInformation>
170   <TextualInformation id="93368" lang="en">
171     <TextSectionList count="0">
172     </TextSectionList>
173   </TextualInformation>
174 </TextualInformationList>
175 </Disorder>

```

LISTING 2: Orphanet XML: Full entry on cross-referencing of cystic fibrosis

The DrugBank entries are even longer, depending on the number of associations with targets, enzymes, transporters or other bioactivities which are listed for a given drug. In Listing 3 a partial entry for Ivacaftor is shown. This is the part of the XML entry from which the target-drug associations in the triangulation effort originate from. For clarity some information has been removed from the XML and replaced by (...). Highlighted in yellow are the name of the drug (line 5), the action of the drug on the target (lines 17-19) and the identifier of the target (line 22).

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <drug created="2012-02-24" type="small molecule" updated="2015-10-29"
   ↪ xmlns="http://www.drugbank.ca"
   ↪ xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
3   <drugbank-id primary="true">DB08820</drugbank-id>
4   <drugbank-id>DB05989</drugbank-id>
5   <name>Ivacaftor</name>
6   <description>Ivacaftor (also known as Kalydeco or VX-770) is a drug for the
   ↪ treatment of cystic fibrosis, developed by Vertex Pharmaceuticals and the
   ↪ Cystic Fibrosis Foundation. FDA approved on January 31, 2012.
   ↪ </description>
7   <cas-number>873054-44-5</cas-number>
8   <groups>
9     <group>approved</group>
10  </groups>
11  ...
12  <targets>
13    <target position="1">
14      <id>BE0001195</id>
15      <name>Cystic fibrosis transmembrane conductance regulator</name>
16      <organism>Human</organism>
17      <actions>
18        <action>potentiator</action>
19      </actions>
20  <references># Yu H, Burton B, Huang CJ, Worley J, Cao D, Johnson JP Jr,
   ↪ Urrutia A, Joubran J, Seepersaud S, Sussky K, Hoffman BJ, Van Goor F:
   ↪ Ivacaftor potentiation of multiple CFTR channels with gating mutations.
   ↪ J Cyst Fibros. 2012 Jan 30.
   ↪ "Pubmed":http://www.ncbi.nlm.nih.gov/pubmed/ 22293084</references>
21  <known-action>yes</known-action>
22  <polypeptide id="P13569" source="Swiss-Prot">
23    <name>Cystic fibrosis transmembrane conductance regulator</name>
24    <general-function>Involved in chloride ion transport</general-function>
25    <specific-function>Involved in the transport of chloride ions. May regulate
   ↪ bicarbonate secretion and salvage in epithelial cells by regulating
   ↪ the SLC4A7 transporter</specific-function>
26    <gene-name>CFTR</gene-name>
27    <locus>7q31.2</locus>

```

```

28     <cellular-location>Membrane; multi-pass membrane
        ↳ protein</cellular-location>
29     <transmembrane-regions>...</transmembrane-regions>
30     <signal-regions>
31     </signal-regions>
32     <theoretical-pi>9.02</theoretical-pi>
33     <molecular-weight>168144.0</molecular-weight>
34     <chromosome-location>
35     </chromosome-location>
36     <organism ncbi-taxonomy-id="9606">Human</organism>
37     <external-identifiers>
38     ...
39     </external-identifiers>
40     <synonyms>
41     <synonym>ATP-binding cassette transporter sub- family C member
        ↳ 7</synonym>
42     <synonym>cAMP- dependent chloride channel</synonym>
43     <synonym>CFTR</synonym>
44     </synonyms>
45     <amino-acid-sequence format="FASTA">...</amino-acid-sequence>
46     <gene-sequence format="FASTA">...</gene-sequence>
47     <pfams>
48     <pfam>
49     <identifier>PF00005</identifier>
50     <name>ABC_tran</name>
51     </pfam>
52     <pfam>
53     <identifier>PF00664</identifier>
54     <name>ABC_membrane</name>
55     </pfam>
56     </pfams>
57     <go-classifiers>
58     ...
59     </go-classifiers>
60     </polypeptide>
61     </target>
62     </targets>
63     <enzymes>
64     ...
65     </enzymes>
66     <carriers>
67     ...
68     </carriers>
69     <transporters>
70     ...
71     </transporters>
72 </drug>

```

LISTING 3: DrugBank XML segment: Targets of Ivacaftor

2.2.2 Data Retrieval: Download, API Calls

As mentioned in the introduction to the KNIME[®] Analytics Platform, the handling of data requires an import into the internal format. This can be done in various ways.

If the user already has a dataset the method of choice are dedicated nodes from the *Node repository* as seen in Figure 10. The *IO* (Input/Output) section provides among others *Reader* nodes for standard file formats.

In my case the data had to be downloaded from the respective sources. The retrieval and import into KNIME[®] are described in more detail in subsubsection 2.2.2.1.

If however the data is not present as a file because the user is accessing databases through web services, then the retrieval can be done directly from within KNIME[®] as described in subsubsection 2.2.2.2.

Alternately, some databases and platforms provide dedicated nodes. The access to the Open PHACTS platform was originally configured with the dedicated Open PHACTS KNIME[®] nodes. These nodes were developed during the Open PHACTS project and can be downloaded from GitHub³³. The nodes however have several drawbacks which require the user to add additional nodes for proper functionality. Therefore in an attempt to reduce the number of nodes used in the workflow and to make the workflow more elegant in general I discarded the Open PHACTS nodes and implemented another solution (a detailed description of the solution is given in section 2.3).

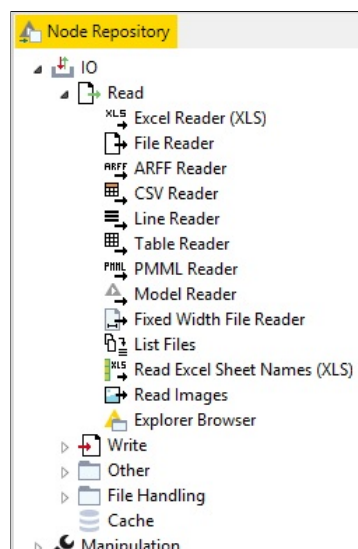


FIGURE 10: Node repository - Import nodes (screenshot)

2.2.2.1 Download

The Orphanet dataset as well as the DrugBank dataset were downloaded from their respective sources as described in section 2.1. It turned out that for the final evaluation of the triangulated links between diseases, targets and drugs specific DrugBank data was needed, which was not included in the Open PHACTS Discovery Platform. Therefore during one of the workflow adjustments I downloaded DrugBank directly and incorporated it into the workflow. The Orphanet as well as the DrugBank dataset were in the XML file format and required the installation of the “KNIME XML-Processing” extension.

The XML nodes are located in the *Structured data* section of the node repository (Figure 11) and provide an *XML Reader* which allows for an import of XML files.

The *XML Reader* needs to be configured (Figure 12) with the path to the file. There is also an option to use an XPath filter to load the XML file in a structured way (depending on the XPath, entries can be loaded in separate cells) instead of loading the entire XML as an entry in a single cell.

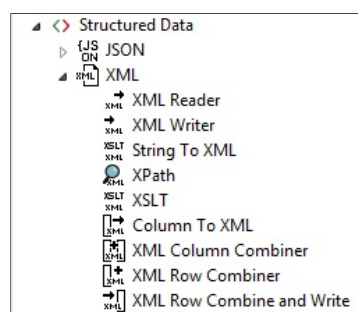


FIGURE 11: Node repository - XML nodes (screenshot)

33 <https://github.com/openphacts/OPS-Knime>, accessed on 19.01.2018

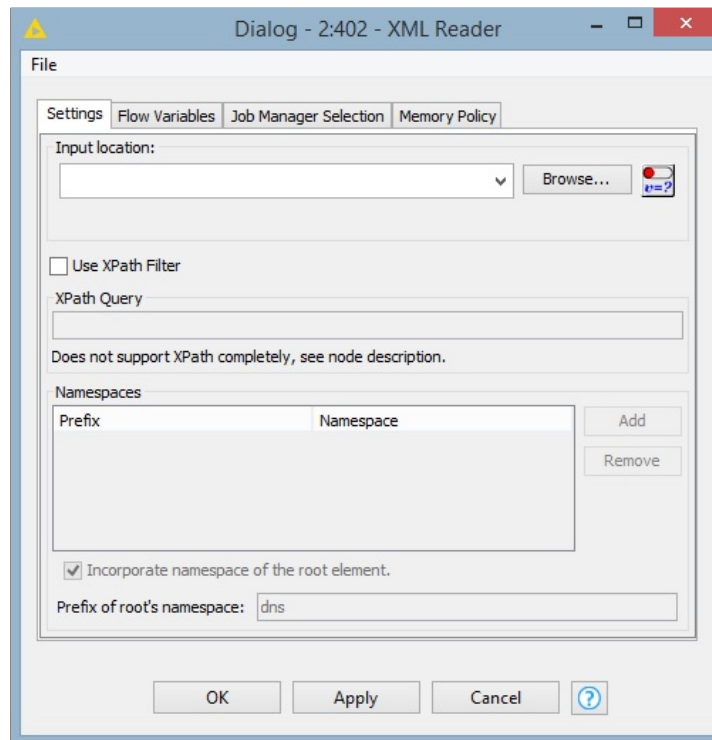


FIGURE 12: XML Reader - configuration (screenshot)

2.2.2.2 API Calls

Some databases provide their datasets through an Application Programming Interface (API), allowing for a retrieval of specific data without having to download the entire database. The API provides a set of standard commands for performing common operations, in the case of databases queries for data. The application then interprets the retrieved data and presents the user with the desired information in a readable format. Often the user can choose between different formats such as XML or JSON.

The set of standard commands are called API calls. Each database that provides an API has therefore a defined set of questions and operations (the calls usually span all queries which are possible when searching the database online and at times API calls provide even additional functionalities) implemented as a certain Uniform Resource Locator (URL)³⁴ syntax which is similar to a website address. This syntax is then converted into the appropriate query language by the API and the results from the query are presented to the user in a readable way.

Some databases provide direct access to their API and others require access keys. These would then have to be added to the query, as is the case when accessing the Open PHACTS API. The user can either sign up³⁵ to be provided with access keys or use the Open PHACTS KNIME[®] nodes which are equipped with them (as seen in Figure 13).

To create an API call first the user has to determine the correct syntax for their question or operation and then adapt it. The API calls for the Open PHACTS

³⁴ <https://en.wikipedia.org/wiki/URL>, accessed on 29.01.2018

³⁵ <https://dev.openphacts.org/signup>, accessed on 29.01.2018

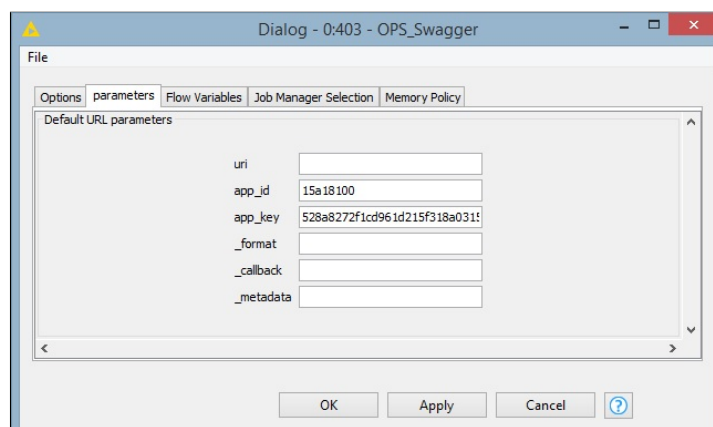


FIGURE 13: OPS_Swagger - access keys (screenshot)

platform³⁶ all start out with the syntax *https://beta.openphacts.org/2.2/*, 2.2 being the current version (at the time of writing) of the Open PHACTS API. The rest depends on the question.

If for example the user wants to retrieve targets based on a disease, the appropriate API call starts out with *https://beta.openphacts.org/2.2/disease/getTargets?*. The question mark indicates the end of the prefix and the beginning of the customisation.

What follows now is a disease specific Uniform Resource Identifier (URI)³⁷. The example disease in the API documentation is atrial fibrillation with the corresponding URI *http://linkedlifedata.com/resource/umls/id/C0004238*. The combination of a letter and numbers at the end of the URI is the UMLS identifier of the disease. Therefore the API call up until now is *https://beta.openphacts.org/2.2/disease/getTargets?uri=http%3A%2F%2Flinkedlifedata.com%2Fresource%2Fumls%2Fid%2FC0004238*. Special characters such as for example / have to be escaped when used in a URL, meaning that they need to be replaced by a specific set of characters (/ is replaced by %2F). This is called URL encoding³⁸.

The next part of the URL are the access keys, meaning the app id and the app key as seen in Figure 13 in the *OPS_Swagger* node. The delimiter between optional parameters of the URL is &. The URL with the prefix, the disease URI and the access keys therefore is *https://beta.openphacts.org/2.2/disease/getTargets?uri=http%3A%2F%2Flinkedlifedata.com%2Fresource%2Fumls%2Fid%2FC0004238&app_id=15a18100&app_key=528a8272f1cd961d215f318a0315dd3d*.

The preceding parameters are all vital. Without the disease identifier or without the access keys results will not be retrieved. All following parameters are however optional. For example Open PHACTS by default retrieves up to 10 results. If the user wishes to retrieve all results in a single page the parameter *_pageSize* has to be set to *all*. The format of the response can be customised by setting the parameter *_format* for example to *xml*.

36 Open PHACTS Foundation. *OpenPHACTS Development Version API Active Docs* Accessed on 29.01.2018. <<https://dev.openphacts.org/docs/2.2>>.

37 https://en.wikipedia.org/wiki/Uniform_Resource_Identifier, accessed on 29.01.2018

38 https://www.w3schools.com/tags/ref_urlencode.asp, accessed on 29.01.2018

For the above mentioned disease the full URL including the last two optional parameters therefore is `https://beta.openphacts.org/2.2/disease/getTargets?uri=http%3A%2F%2Flinkedlifedata.com%2Fresource%2Fumls%2Fid%2FC0004238&app_id=15a18100&app_key=528a8272f1cd961d215f318a0315dd3d&_pageSize=all&_format=xml`.

In KNIME[®] this syntax can be created using for example the *String Manipulation* node. Afterwards the resource can be retrieved with the *GET Request* node (the ‘Timeout (s)’ option within the node configuration is by default very short and can be set to for example ‘300’ to increase the likelihood of a successful retrieval of the XML file). When typed into a browser the result is in XML format (Figure 14), describing the genes (in blue rectangles) associated with the disease. The XML can then be mined with an *XPath* node.



```

<?xml version="1.0" encoding="UTF-8"?>
<result format="linked-data-api" version="2.2" href="http://beta.openphacts.org/3002/2.2/disease/getTargets?uri=http%3A%2F%2Flinkedlifedata.com%2Fresource%2Fumls%2Fid%2FC0004238&app_id=80c9eefe&app_key=b056a66b85925e0c4cb497fa8b6bcb24&_pageSize=all&_format=xml&_page=1">
  <label>Targets for Disease: List</label>
  <type href="http://purl.org/linked-data/api/vocab#Page"/>
  <items>
    <item href="http://identifiers.org/ncbigene/10008">
      <forDisease href="http://linkedlifedata.com/resource/umls/id/C0004238">
        <name>Atrial Fibrillation</name>
        <inDataset href="http://rdf.disgenet.org/v2.1.0/void.ttl#disease"/>
      </forDisease>
      <inDataset href="http://rdf.disgenet.org/v2.1.0/void.ttl#gene"/>
    </item>
    <item href="http://identifiers.org/ncbigene/10021">
      <forDisease href="http://linkedlifedata.com/resource/umls/id/C0004238">
        <name>Atrial Fibrillation</name>
        <inDataset href="http://rdf.disgenet.org/v2.1.0/void.ttl#disease"/>
      </forDisease>
      <inDataset href="http://rdf.disgenet.org/v2.1.0/void.ttl#gene"/>
    </item>
    <item href="http://identifiers.org/ncbigene/100271279">
      <forDisease href="http://linkedlifedata.com/resource/umls/id/C0004238">
        <name>Atrial Fibrillation</name>
        <inDataset href="http://rdf.disgenet.org/v2.1.0/void.ttl#disease"/>
      </forDisease>
    </item>
  </items>
</result>

```

FIGURE 14: Disease URL - browser result (screenshot)

The approach for the generation of API calls described above may be shortened by using the dedicated Open PHACTS KNIME[®] nodes. For detailed information on the use of the dedicated OPEN PHACTS KNIME[®] nodes consult the book chapter ‘Accessing the Open PHACTS Discovery Platform with Workflow Tools’ by Digles *et al.* [73]. The use of these nodes however requires several additional workflow steps. First of all the *OPS_Swagger* (the node responsible for the generation of API calls) can only be configured to retrieve calls for the API version 1.4 and 1.5 (Figure 15). To access the current version (2.2) based on the *OPS_Swagger* node the user has to add a *String Replacer* or a *String Manipulation* node and exchange the version in the URL manually.

The second requirement is that the user filters out all columns preceding the *OPS_Swagger* except for the column with the input data. The name of this column also has to match the name of the URL parameter exactly, which sometimes requires additional renaming of the column.

The retrieval has to be configured as a loop because the *OPS_Swagger* processes natively only the first row of data (which is usually one of many) and simultaneously removes all used columns. Therefore if the user wants to keep the preceding information, nodes for variables have to be introduced to first save the

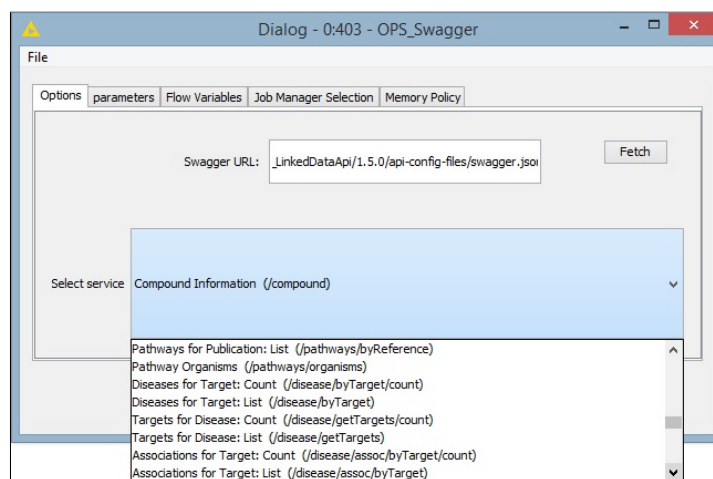


FIGURE 15: OPS_Swagger - API (screenshot)

preceding data and to then add it again to the table right before the loop ends. Additionally, the *OPS_Swagger* only generates the URL but does not retrieve the data, which is why additional nodes, such as the *GET Request* node, have to be added.

These requirements often result in many additional nodes for data retrieval through Open PHACTS which is why I prefer to look up the syntax in the Open PHACTS API documentation and then to generate it with a single *String Manipulation* node. This way the data retrieval can be reduced to two nodes (*String Manipulation* and *GET Request*) which results in a compact workflow.

2.2.3 Mining of textual data

Text mining is a method for the extraction of information from text. The method can be employed in various contexts and is of special interest in the field of life sciences where a majority of insights and results from research are released to the public in the form of published papers. These texts contain valuable information which cannot be mined directly because a text by nature is not a database. These texts have to be structured first to derive patterns upon which an evaluation and interpretation can be done. Text mining usually consists of various steps such as parsing, tagging, filtering and many more, all of which are used to structure text in a machine readable way (such as a document vector where the absence or presence of words in a document is designated by 0 or 1) so that a further analysis can be performed.

One of the simplest examples of text mining is a word count which is then visualised as a word cloud (Figure 16).

For text mining in KNIME[®] an extension called *KNIME Textprocessing* is available. Example workflows are available on the KNIME[®] EXAMPLES Server³⁹ and based on them I created a solution which fit the needs of my workflow. The extension provides a variety of nodes which can be used for text mining tasks. An

³⁹ Example workflows documentation, <https://www.knime.com/example-workflows>, accessed in May 2018

can be added directly after that (as shown in Figure 16) or the results can be analysed further. Such an analysis might be a hierarchical clustering, a classification or any other machine learning method for which there are nodes available.

In the workflow abstracts were retrieved through the EMBL-EBI web services which are described in section 2.1.7. Thusly retrieved abstracts were then mined in an effort to categorise the disease-gene relationship and to extract meaningful associations from the triangulation results. The aim of the mining effort was to categorise abstracts, which can be done if a training set with assigned categories is present. Such a training set consists of abstracts to which a category (e.g. upregulation) has been assigned. Based on these assigned categories and their corresponding structured text, categories can be assigned to texts with unknown categories. This works quite well in fairly easy cases as described in several KNIME® tutorials [74–76]. A limiting factor in text mining efforts is the complexity of the categorisation. As long as the possible categories are directly related to words in the text chances are the categorisation will deliver satisfactory results. For example if the categorisation effort needs to differentiate between German and Italian restaurants based on reviews some form of the word German or Italian will most likely be found in the review. Therefore it will be easy for a classifier to determine the correct category. However for complex categories the solution needs to be meticulously customised.

Due to technical reasons (text mining of thousands of abstracts requires a lot of computational power) as well as the nature of the task (without a training set a categorisation does not produce meaningful results, and the compilation of an excellent training set requires time) the text mining part of the workflow is currently incomplete in regard to results.

2.3 THE WORKFLOW

The workflow consists of three major steps, an intermediate step and the analysis of the results. In the three major steps diseases, genes and drugs are extracted, in the intermediate step text mining is performed on abstracts of publications for a better understanding of the disease-gene relationship, and in the final step the results of the workflow are evaluated. A schematic depiction of the steps in the workflow is shown in Figure 18.

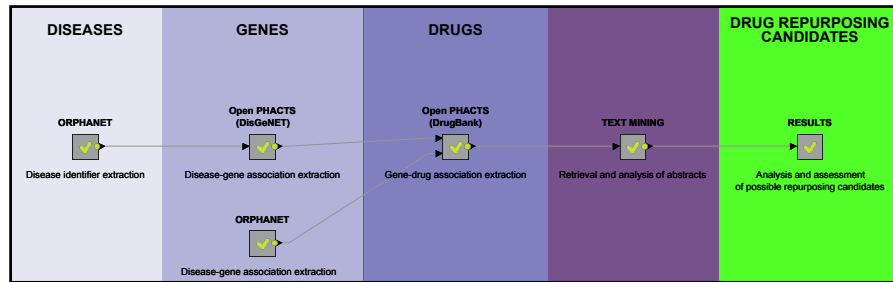


FIGURE 18: Schematic view of the workflow

The following sections elaborate on essential steps of the data mining process and do not mention each node and the corresponding configuration that was used. The choice to elaborate on essential steps only was made partially because the workflow consists of over 60 nodes, some of which are trivial, for example those that perform simple joining tasks. More importantly the choice was made because the workflow does not require a fixed configuration of nodes, which is due to the nature of KNIME[®] where most tasks can be executed in several ways. In fact I myself have devised multiple versions of the workflow, especially more efficient ones the more familiar I became with KNIME[®], but also ones that are for example better suited for a certain type of analysis or ones that perform one of the steps in higher or less detail.

2.3.1 Extraction of diseases

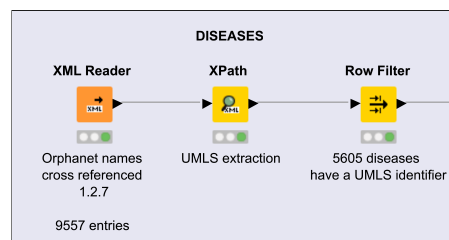


FIGURE 19: Nodes for the extraction of UMLS identifiers of orphan diseases

The workflow starts out by reading in the ‘Rare diseases and cross-referencing’⁴¹ dataset, an XML file which includes cross-references between Orphanet identifiers and identifiers from other databases. An XML entry from this dataset for cystic fibrosis is shown in Listing 2 on page 36. The *XML Reader* provides the

⁴¹ Orphadata. *Rare diseases and cross-referencing* Accessed on 31.10.2018. <http://www.orphadata.org/data/xml/en_product1.xml>.

option of an XPath filter and for this XML file the XPath filter `//DBOR/Disorder-List/Disorder` was applied to divide the file into separate disease entries, shown in the KNIME[®] table as one disease per row as seen in Figure 20. The prefix of the root's namespace is also incorporated and specified as 'dns', which is a configuration I took over from the KNIME[®] Example Workflow '01203001_drugBank'.

Row ID	XML XML
Row0	<pre><?xml version="1.0" encoding="UTF-8"?> <Disorder id="17601"> <OrphaNumber>166024</OrphaNumber> <ExpertLink lang="en">http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=eng&Expert=166024</ExpertLink> <Name lang="en">Multiple epiphyseal dysplasia, Al-Gazali type</Name> </Disorder></pre>
Row1	<pre><?xml version="1.0" encoding="UTF-8"?> <Disorder id="17603"> <OrphaNumber>166032</OrphaNumber> <ExpertLink lang="en">http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=eng&Expert=166032</ExpertLink> <Name lang="en">Multiple epiphyseal dysplasia, with miniepiphyses</Name> </Disorder></pre>
Row2	<pre><?xml version="1.0" encoding="UTF-8"?> <Disorder id="2"> <OrphaNumber>58</OrphaNumber> <ExpertLink lang="en">http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=eng&Expert=58</ExpertLink> <Name lang="en">Alexander disease</Name> </Disorder></pre>

FIGURE 20: 'Rare diseases and cross-referencing' dataset as seen in the KNIME[®] table

The UMLS identifiers relevant for the connection to Open PHACTS are retrieved with an XPath node. The XPath expression for the retrieval is `/Disorder/ExternalReferenceList/ExternalReference/Source[.="UMLS"]/following-sibling::Reference`. The XPath has this specific configuration because the XML file includes many different identifiers (e.g. OMIM in line 25, MeSH in line 35), therefore the XPath has to specify the `<Source>` as UMLS (in line 49) and the desired result as the entry that immediately follows the `<Source>` because that entry is the desired identifier. The `<Source>` and `<Reference>` entries are at the same level, they are therefore siblings and the `<Reference>` entry can be called with the syntax `following-sibling::Reference`. Additional information such as the name of the disease or its Orphanet identifier are also retrieved, these XPaths however have a simpler syntax than the one mentioned above and can be created easily in the configuration window of the XPath node by clicking on the desired entry. Lastly an important configuration of the XPath node is to set the 'Multiple tags options' of the XPath setting to 'Multiple Rows' whenever there might be multiple values for the desired entry. Figure 21 illustrates such cases. Each disease listed here has two different UMLS identifiers which can only be retrieved by the above mentioned setting. The default KNIME[®] setting is 'Single Cell' which would result in a single row.

Row ID	S Orpha Name	S UMLS	S MeSH	S OMIM	S Orpha...
Row7_1	Aspartylglucosaminuria	C0268225	C538402	208400	93
Row7_2	Aspartylglucosaminuria	C2931840	D054880	?	93
Row8_1	Multiple sulfatase deficiency	C0268263	?	272200	585
Row8_2	Multiple sulfatase deficiency	C1720864	?	?	585
Row9_1	Beta-mannosidosis	C0342849	D044905	248510	118
Row9_2	Beta-mannosidosis	C2931893	?	?	118
Row10_1	Canavan disease	C0206307	D017825	271900	141
Row10_2	Canavan disease	C3542499	?	?	141

FIGURE 21: Results of the identifier extraction as seen in the KNIME[®] table

The resulting table is then filtered for entries which have at least one UMLS identifier. The filtering concludes the first major step of the workflow where UMLS identifiers are retrieved for a subsequent connection to the Open PHACTS platform.

2.3.2 Extraction of genes

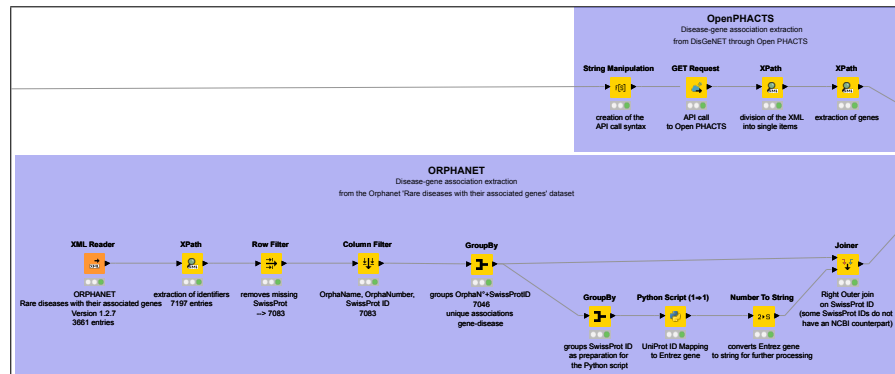


FIGURE 22: Nodes for the retrieval of genes through Open PHACTS based on UMLS identifiers and for the extraction of genes from the 'Rare diseases with their associated genes' dataset from Orphadata

Genes for diseases are retrieved in two subsections of the workflow from two different sources. The first subsection, labeled 'Open PHACTS' in Figure 22, retrieves genes through API calls to the Open PHACTS platform, based on the UMLS identifiers which were retrieved in the first part of the workflow. The second subsection, labeled 'ORPHANET' in Figure 22, mines the 'Rare diseases with their associated genes' dataset⁴², an XML file provided by Orphadata with links between diseases and their genes.

The syntax of the API calls is generated in a *String Manipulation* node based on the expression `join("https://beta.openphacts.org/2.2/disease/assoc/byDisease?uri=http%3A%2F%2Flinkedlifedata.com%2Fresource%2Fumls%2Fid%2F", $UMLS$, "&app_id=15a18100&app_key=528a8272f1cd961d215f318a0315dd3d&_pageSize=all&_format=xml")`. This expression specifies that the prefix for the 'Associations for Disease' API call shall be joined with a given UMLS identifier (denoted by `$UMLS$` between two commas, meaning that the value is taken directly from the column named `UMLS` from the KNIME[®] table) and further parameters such as the API keys, the size of the pages that are to be returned and the desired format of the response. For example in case of cystic fibrosis and its UMLS identifier `C0010674` the corresponding API call is `https://beta.openphacts.org/2.2/disease/assoc/byDisease?uri=http%3A%2F%2Flinkedlifedata.com%2Fresource%2Fumls%2Fid%2FC0010674&app_id=15a18100&app_key=528a8272f1cd961d215f318a0315dd3d&_pageSize=all&_format=xml`. The API calls are then executed for the entire list of diseases with a *GET Request* node. The resulting XML files are used for the extraction of relevant entries. An example of such an XML file of cystic fibrosis and its association with the phosphodiesterase 5A (National Center for Biotechnology Information (NCBI) gene 8654) based on the publication 'Sildenafil (Viagra) corrects Δ F508-CFTR location in nasal epithelial cells from patients with cystic fibrosis' [77] is shown in Listing 4.

42 Orphadata. *Rare diseases with their associated genes* Accessed on 31.10.2018. <http://www.orphadata.org/data/xml/en_product6.xml>.

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <item href="http://rdf.disgenet.org/gene-disease-association.ttl#DGN6674c21b7b2f4_
  ↪ d015ad6c300179097f4">
3   <type href="http://semanticscience.org/resource/SIO_001121">
4     <label>gene-disease biomarker association</label>
5   </type>
6   <primarySource href="http://rdf.disgenet.org/v2.1.0/void.ttl#befree-20140225">
7   </primarySource>
8   <pmid href="http://identifiers.org/pubmed/15618584">
9   </pmid>
10  <gene href="http://identifiers.org/ncbigene/8654">
11    <inDataset href="http://rdf.disgenet.org/v2.1.0/void.ttl#gene">
12    </inDataset>
13  </gene>
14  <disease href="http://linkedlifedata.com/resource/umls/id/C0010674">
15    <name>Cystic Fibrosis</name>
16    <diseaseClass>
17      <item href="http://rdf.imim.es/rh-mesh.owl#C06">
18        <name>Digestive System Diseases</name>
19        <inDataset
20          ↪ href="http://rdf.disgenet.org/v2.1.0/void.ttl#diseaseClass">
21        </inDataset>
22      </item>
23      <item href="http://rdf.imim.es/rh-mesh.owl#C08">
24        <name>Respiratory Tract Diseases</name>
25        <inDataset
26          ↪ href="http://rdf.disgenet.org/v2.1.0/void.ttl#diseaseClass">
27        </inDataset>
28      </item>
29      <item href="http://rdf.imim.es/rh-mesh.owl#C16">
30        <name>Congenital, Hereditary, and Neonatal Diseases and
31          ↪ Abnormalities</name>
32        <inDataset
33          ↪ href="http://rdf.disgenet.org/v2.1.0/void.ttl#diseaseClass">
34        </inDataset>
35      </item>
36    </diseaseClass>
37    <inDataset href="http://rdf.disgenet.org/v2.1.0/void.ttl#disease">
38    </inDataset>
39  </disease>
40  <inDataset
41    ↪ href="http://rdf.disgenet.org/v2.1.0/void.ttl#geneDiseaseAssociation">
42  </inDataset>
43  <description>[The findings provide proof of principle for sildenafil as a
44    ↪ DeltaF508-CFTR trafficking drug and give encouragement for future testing
45    ↪ of sildenafil and related PDE5 inhibitors in patients with CF.]. Sentence
46    ↪ from MEDLINE/PubMed, a database of the U.S. National Library of
47    ↪ Medicine.</description>
48 </item>

```

LISTING 4: Example XML resulting from an 'Associations for Disease' API call

Several entries in the XML file are of interest for further processing, such as the *pmid* entry in line 8, the PubMed identifier of the publication the association between the disease and the gene is based on (which will be further used in the *Text Mining* part of the workflow), or the *gene* entry in line 10 which denotes the specific gene the disease is associated with (which will be used for the connection to drugs). Both identifiers are later processed in *String Manipulation* nodes and this is one of the cases where KNIME® offers multiple solutions for one task.

An easy way is to extract here the entire URL with an XPath which retrieves the attribute *@href* (this can be done visually by clicking on the desired attribute

in the XPath configuration and by adding the proposed XPath) and then for example to remove the prefix (e.g. 'http://identifiers.org/pubmed/') with a *String Replacer* node.

A more efficient way however is to compile an XPath which extracts only the number. This requires proficiency in XPath expressions and cannot be done visually within the node configuration. The advantage in this case is that the workflow is quite elegant (there is only one node instead of two) and it is also more efficient. Such an XPath utilizes the *substring-after* functionality to retrieve only the substring after a given prefix. For the *pmid* the expression is *substring-after(//pmid/@href,'http://identifiers.org/pubmed/')*. It specifies that the node *pmid* at any level within the tree shall be addressed (this is done with the use of // and is a precaution in case the node does not always appear at the same position within different XML files so that whenever a *pmid* entry is present the information is retrieved, instead of an incomplete retrieval where only entries at the exact same position are retrieved). The expression further specifies that the attribute *@href* shall be addressed and that the part of the attribute that shall be retrieved is the substring after the prefix *http://identifiers.org/pubmed/*. For the gene entry the only difference is the last part where the *substring-after* functionality is used on the prefix *http://identifiers.org/ncbigene/*. This way the workflow is able to directly retrieve the identifiers that are used for further processing.

The second part of the retrieval of disease-gene associations, which is based on Orphanet data, starts out by reading in the 'Rare diseases with their associated genes' dataset⁴³ with an *XML Reader*. The XPath filter *//DBOR/DisorderList/Disorder* is used to divide the XML into single disease entries per row, and as described in subsection 2.3.1 for the 'Rare diseases and cross-referencing dataset' the prefix of the root's namespace is again specified as *dns*. Afterwards information on the diseases and their genes is extracted with an *XPath* node. The XML file includes several sources of alternative gene identifiers such as HGNC⁴⁴, SwissProt⁴⁵ or OMIM⁴⁶.

Given that the information will have to be combined with the results from Open PHACTS which uses NCBI identifiers⁴⁷, the identifiers from both datasets have to be unified. This can again be done in different ways. Open PHACTS for example provides a *Mapping service between URLs* for URLs of identifiers contained within the platform and UniProt⁴⁸ provides a mapping service between identifiers from a number of databases. Given that SwissProt identifiers are present in the dataset and UniProt provides a mapping service dedicated to its identifiers the workflow utilizes this service instead of the one provided by Open PHACTS. In preparation for the mapping the dataset is filtered to include only entries that have SwissProt identifiers and the table is reduced to relevant columns only with

43 Orphadata. *Rare diseases with their associated genes* Accessed on 31.10.2018. <http://www.orphadata.org/data/xml/en_product6.xml>.

44 HUGO Gene Nomenclature Committee, <https://www.genenames.org/>, accessed on 07.05.2018

45 Broadly speaking SwissProt is the manually annotated and reviewed part of the UniProt (Universal Protein Resource) database. See also <http://www.uniprot.org/help/about>, accessed on 07.05.2018

46 Online Mendelian Inheritance in Man, <https://www.omim.org/>, accessed on 07.05.2018

47 <https://www.ncbi.nlm.nih.gov/>, accessed on 07.05.2018

48 http://www.uniprot.org/help/api_idmapping, accessed on 07.05.2018

the *Column Filter*. To ensure that each association between a disease and a gene is unique a *GroupBy* node is additionally employed.

The mapping service is executed with a Python script⁴⁹ provided by UniProt which I adapted for the use in KNIME[®] (shown in Listing 5) and which requires the input to be specifically formatted. The required format is generated with a *GroupBy* node which aggregates unique UniProt accession numbers by concatenating them with a comma in between, resulting in a single cell with all identifiers that need to be mapped (e.g. 'Q13315, P05067, P01034, P18564, Q13751, Q9BZG2, Q9NP70, ...'). This *GroupBy* node feeds into a *Python Script (1→1)* node, which is configured with the script shown in Listing 5.

```

1 import urllib,urllib2,pandas
2 from StringIO import StringIO
3
4 url = 'http://www.uniprot.org/uploadlists/'
5
6 params = {
7     'from':'ID',
8     'to':'P_ENTREZGENEID',
9     'format':'tab',
10    'query':input_table['SwissProt ID'].iloc[0]
11 }
12
13 data = urllib.urlencode(params)
14 request = urllib2.Request(url, data)
15 contact = "" # Please set your email address here to help us debug in case of
16 ↪ problems.
17 request.add_header('User-Agent', 'Python %s' % contact)
18 response = urllib2.urlopen(request)
19 page = response.read()
20
21 output_table = pandas.read_csv(StringIO(page), sep="\t")
22
23 output_table.columns = ['UniProt ID', 'NCBI']

```

LISTING 5: Python script for the mapping of UniProt and NCBI identifiers

The script specifies that the conversion shall be done from UniProt IDs (line 7) to NCBI IDs (line 8) based on the input from the KNIME[®] table, specifically the column named 'SwissProt ID' and its first cell (line 10). The last line of the script specifies that after the conversion and retrieval of the identifiers the columns shall be renamed to 'UniProt ID' and 'NCBI' to be easily recognizable afterwards. The data as seen in the KNIME[®] table before and after the conversion is shown in Figure 23.

Now that both the top and the bottom section of the disease-gene association extraction (as seen in Figure 22 on page 47) are equipped with NCBI identifiers both sources can be joined into one table which will be used for the retrieval of drugs and publication abstracts.

⁴⁹ Programmatic access - Mapping database identifiers, https://www.uniprot.org/help/api_idmapping, accessed on 07.05.2018

Row ID	S OrphaN...	S SwissProt ID	S Orpha Name
Row0	100	Q13315	Ataxia-telangiectasia
Row1	100006	P05067	ABeta amyloidosis, Dutch type
Row2	100008	P01034	ACys amyloidosis
Row3	100031	P18564	Hypoplastic amelogenesis imperfecta
Row4	100031	Q13751	Hypoplastic amelogenesis imperfecta

(A) Table as seen before identifier mapping

Row ID	S OrphaN...	S Orpha Name	S UniProt ID	S NCBI
Row0_0	100	Ataxia-telangiectasia	Q13315	472
Row1_1	100006	ABeta amyloidosis, Dutch type	P05067	351
Row2_2	100008	ACys amyloidosis	P01034	1471
Row3_3	100031	Hypoplastic amelogenesis imperfecta	P18564	3694
Row4_4	100031	Hypoplastic amelogenesis imperfecta	Q13751	3914

(B) Table as seen after identifier mapping

FIGURE 23: Disease-gene associations from Orphanet before and after identifier mapping

2.3.3 Extraction of drugs

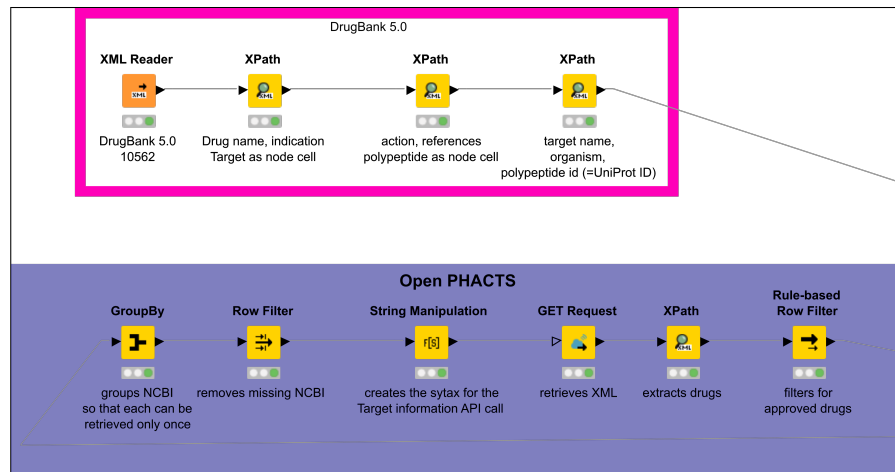


FIGURE 24: Nodes for the retrieval of drugs based on genes through Open PHACTS and the addition of annotations directly from DrugBank

This part of the workflow starts out with the list of gene identifiers which resulted from the joining of data retrieved through Open PHACTS and from Orphanet, as described in the previous section. The current dataset includes multiple entries for a gene whenever multiple publications were retrieved for a certain disease-gene association, which is why this part starts out with a *GroupBy* node. The node groups the gene identifiers to ensure that every identifier is unique. This is done to reduce the computational strain on the workflow, because this way for each of the genes a single API call is made, regardless of the number of publications or diseases associated with it. Afterwards the information is again connected to diseases associated with the gene. Without this step around 330,000 API calls would have to be made, whereas when using this intermediate step only around 12,000 are needed. API calls take a lot of time whereas the rejoining of the information is much faster.

The *Row Filter* ensures that only rows with valid NCBI identifiers are taken into account (otherwise the *GET Request* node results in an error).

The preparation for the third major step of the workflow is now finished, what follows now is the creation of the syntax for the *Target Information* API calls which will retrieve information on drugs associated with targets (protein products of genes). The expression within the *String Manipulation* node - `join("https://beta.openphacts.org/2.1/target?uri=http%3A%2F%2Fidentifiers.org%2Fncbigene%2F", $NCBI$, "&app_id=15a18100&app_key=528a8272f1cd961d215f318a0315dd3d&format=xml")` - is similar to the one from the last step, it is again a joining of a prefix, the identifier taken from the KNIME[®] table and a suffix with parameters. For the phosphodiesterase 5A, the gene mentioned in connection to cystic fibrosis in one of the previous paragraphs, the API call is therefore `https://beta.openphacts.org/2.1/target?uri=http%3A%2F%2Fidentifiers.org%2Fncbigene%2F8654&app_id=15a18100&app_key=528a8272f1cd961d215f318a0315dd3d&format=xml`. It is important to know that at this point information is retrieved only if an entry for the respective gene is present in the Open PHACTS database. As the current dataset is a combination of data from Open PHACTS and Orphanet, entries for some of the genes may not be present.

Afterwards the drug entries are extracted with an *XPath* node and filtered to include approved drugs only. A simplified version (references to publications were for example removed) of the table resulting from this step is shown in Figure 25(A).

Row ID	\$ Drug name	\$ Overall target name	\$ polype...	\$ Orpha Name
Row0	Acebutolol	Beta-2 adrenergic receptor	P07550	Cystic fibrosis
Row1	Acetaminophen	Prostaglandin G/H synthase 2	P35354	Cystic fibrosis
Row2	Acetylsalicylic acid	Endothelin-1 receptor	P25101	Cystic fibrosis
Row3	Acetylsalicylic acid	Prostaglandin G/H synthase 2	P35354	Cystic fibrosis
Row4	Adalimumab	Tumor necrosis factor	P01375	Cystic fibrosis
Row5	Adenine	Adenine phosphoribosyltransferase	P07741	Cystic fibrosis
Row6	Adenosine monophosphate	5'-AMP-activated protein kinase catalytic subunit alpha-1	Q13131	Cystic fibrosis
Row7	Adenosine monophosphate	5'-AMP-activated protein kinase subunit beta-1	Q91478	Cystic fibrosis
Row8	Aflibercept	Vascular endothelial growth factor A	P15692	Cystic fibrosis

(A) Table as seen before additional annotation

Row ID	\$ Drug name	\$ Overall target name	\$ action	\$ polype...	\$ Orpha Name
Row0	Acebutolol	Beta-2 adrenergic receptor	partial agonist	P07550	Cystic fibrosis
Row1	Acetaminophen	Prostaglandin G/H synthase 2	inhibitor	P35354	Cystic fibrosis
Row2	Acetylsalicylic acid	Endothelin-1 receptor	?	P25101	Cystic fibrosis
Row3	Acetylsalicylic acid	Prostaglandin G/H synthase 2	inhibitor	P35354	Cystic fibrosis
Row4	Adalimumab	Tumor necrosis factor	antibody	P01375	Cystic fibrosis
Row5	Adenine	Adenine phosphoribosyltransferase	?	P07741	Cystic fibrosis
Row6	Adenosine monophosphate	5'-AMP-activated protein kinase catalytic subunit alpha-1	activator	Q13131	Cystic fibrosis
Row7	Adenosine monophosphate	5'-AMP-activated protein kinase subunit beta-1	activator	Q91478	Cystic fibrosis
Row8	Aflibercept	Vascular endothelial growth factor A	binder	P15692	Cystic fibrosis

(B) Table as seen after additional annotation

FIGURE 25: Gene-drug associations retrieved through Open PHACTS before (A) and after (B) an additional annotation with DrugBank data

At this point the KNIME[®] table includes the disease which was the starting point of the workflow (seen in (A) and (B) in Figure 25 in the columns ‘Orpha Name’), the associated genes/targets and their identifiers (columns ‘Overall target name’ and ‘polype...’ which stands for polypeptide ID) and the name of the drug (in the column ‘Drug name’) that was retrieved based on the mining process. This drug may or may not be a possible drug repurposing candidate, currently this is impossible to determine because there is no information on its mechanism of action. Given that this type of information is not included in the XML that

is retrieved through Open PHACTS (see Listing 4 on page 48) I adapted the workflow to enrich the table with these annotations directly from DrugBank, resulting in the table shown in 25(B) which has an additional column named ‘action’. The part of the workflow where DrugBank data is mined and subsequently connected to the table is in the top half of the Figure 24.

The DrugBank XML is imported into KNIME[®] with an *XML Reader* which is configured to use the XPath filter `/dns:drugbank/dns:drug` and the prefix of the root’s namespace is set to `dns`. This configuration results in one drug XML entry per row in the KNIME[®] table, as shown in Figure 26. The individual drug entries are extensive and nested because of the wealth of information they contain. To correctly extract all of the desired information the workflow employs three XPath nodes.

Row ID	XML XML
Row0	<pre><?xml version="1.0" encoding="UTF-8"?> <drug created="2005-06-18" type="biotech" updated="2017-11-06" xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <drugbank-id primary="true">DB00001</drugbank-id> <drugbank-id>B1D000024</drugbank-id> <drugbank-id>B1D000024</drugbank-id> <name>Leprudin</name> <description>Leprudin is identical to natural hirudin except for substitution of leucine for isoleucine at the N-terminal end of the molecule. </description> </drug></pre>
Row1	<pre><?xml version="1.0" encoding="UTF-8"?> <drug created="2005-06-18" type="biotech" updated="2017-12-18" xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <drugbank-id primary="true">DB00002</drugbank-id> <drugbank-id>B1D000071</drugbank-id> <drugbank-id>B1D000071</drugbank-id> <name>Cetuximab</name> <description>Cetuximab is an epidermal growth factor receptor binding FAb. Cetuximab is composed of the Fv (variable; antigen-binding) and Fc (constant) regions of a human IgG1 antibody. </description> </drug></pre>

(A) Table as seen after the import of the DrugBank XML

Row ID	XML target	Drug name	indication	drugbank-id
Row0_1	<pre><?xml version="1.0" encoding="UTF-8"?> <target xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <id>B1D000048</id> <name>Prothrombin</name> <organism>Human</organism> </target></pre>	Leprudin	For the treatment of heparin-induced thrombocytopenia	DB00001
Row1_1	<pre><?xml version="1.0" encoding="UTF-8"?> <target position="1" xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <id>B1D000076</id> <name>Epidermal growth factor receptor</name> <organism>Human</organism> </target></pre>	Cetuximab	Cetuximab, used in combination with irinotecan, is indicated for the treatment of metastatic colorectal cancer.	DB00002
Row1_2	<pre><?xml version="1.0" encoding="UTF-8"?> <target position="2" xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <id>B1D000090</id> <name>Low affinity immunoglobulin gamma Fc region receptor III-B</name> <organism>Human</organism> </target></pre>	Cetuximab	Cetuximab, used in combination with irinotecan, is indicated for the treatment of metastatic colorectal cancer.	DB00002

(B) Table as seen after the first XPath node

Row ID	Drug name	indication	drugbank-id	Overall target name	Overall target	action	XML polypeptide
Row0_1_1	Leprudin	For the treatment of heparin-induced thrombocytopenia	DB00001	Prothrombin	Human	inhibitor	<pre><?xml version="1.0" encoding="UTF-8"?> <polypeptide id="P00734" source="Swiss-Prot" xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <name>Prothrombin</name> <general-function>Thrombolytic activity</general-function> <specific-function>Thrombin, which cleaves bonds after Arg and Lys, converts fibrinogen to fibrin. </specific-function> </polypeptide></pre>
Row1_1_1	Cetuximab	Cetuximab, used in combination with irinotecan, is indicated for the treatment of metastatic colorectal cancer.	DB00002	Epidermal growth factor receptor	Human	antagonist	<pre><?xml version="1.0" encoding="UTF-8"?> <polypeptide id="P01833" source="Swiss-Prot" xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <name>Epidermal growth factor receptor</name> <general-function>Oligonucleotide binding</general-function> <specific-function>Receptor tyrosine kinase binding ligands of the EGF family. </specific-function> </polypeptide></pre>
Row1_2_1	Cetuximab	Cetuximab, used in combination with irinotecan, is indicated for the treatment of metastatic colorectal cancer.	DB00002	Low affinity immunoglobulin gamma Fc region receptor III-B	Human	antagonist	<pre><?xml version="1.0" encoding="UTF-8"?> <polypeptide id="O79015" source="Swiss-Prot" xmlns="http://www.drugbank.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.drugbank.ca drugbank.xsd"> <name>Low affinity immunoglobulin gamma Fc region receptor III-B</name> <general-function> </general-function> </polypeptide></pre>

(C) Table as seen after the second XPath node

Row ID	Drug name	indication	drugbank-id	Overall target name	Overall target	action	Target ...	polypeptide ...	organism	references
Row0_1_1_1	Leprudin	For the treatment of heparin-induced thrombocytopenia	DB00001	Prothrombin	Human	inhibitor	Prothrombin	P00734	Human	10505537urpe AG: Anticoagulants in acute coronary syndrome. Eur Heart J. 2007;28(18):2181-2188.
Row1_1_1_1	Cetuximab	Cetuximab, used in combination with irinotecan, is indicated for the treatment of metastatic colorectal cancer.	DB00002	Epidermal growth factor receptor	Human	antagonist	Epidermal growth factor receptor	P01833	Human	10480577Kobayashi N, Yamamoto S, Uehara Y, Hata Y, et al. (2007) Cetuximab: a novel anti-epidermal growth factor receptor antibody. J Clin Oncol. 2007;25(18):2481-2488.
Row1_2_1_1	Cetuximab	Cetuximab, used in combination with irinotecan, is indicated for the treatment of metastatic colorectal cancer.	DB00002	Low affinity immunoglobulin gamma Fc region receptor III-B	Human	antagonist	Low affinity immunoglobulin gamma Fc region receptor III-B	O79015	Human	16336752Nyder LC, Aftabizadeh S, Weiner LM, O'Neil CL, et al. (2007) The structure of the human low-affinity IgG Fc receptor (CD32) and its interaction with IgG. Nat Struct Mol Biol. 2007;14(11):1054-1062.

(D) Table as seen after the third XPath node

FIGURE 26: A step by step extraction of annotations from the DrugBank XML as seen in the various KNIME[®] tables

The first one extracts the name of the drug, its indication and its identifier directly. The parts of the XML where proteins interacting with the drug are referenced are extracted as node cells by the XPath node, meaning that the XML is divided further into smaller XML entries, as seen in Figure 26(B). The drug leprudin has only one target and therefore even after a division of the XML results in one row only. Cetuximab however has 12 targets referenced in its Drug-

Bank entry and therefore results in multiple rows after a division, two of which are directly below lepirudin.

Now that the entries have been properly divided the corresponding mechanism of action of the drug on the target can be retrieved. The resulting table is shown in Figure 26(c). It includes the additional column named ‘action’ which denotes whether the drug is for example an inhibitor. Whenever a mechanism of action is not given in the XML, the table will show a question mark.

The last node extracts the protein identifiers and leads to the final table, as shown in Figure 26(d). This information is connected to the table which results from the Open PHACTS retrieval (shown in Figure 25(A)) to form the table shown in Figure 25(B) which brings the workflow a step closer to being able to determine the usefulness of the proposed drug repurposing candidates.

One of the results within this dataset is the connection between cystic fibrosis and the drug vardenafil. As mentioned in subsection 2.3.2, one of the targets that were retrieved for cystic fibrosis is the phosphodiesterase 5A with the corresponding publication ‘Sildenafil (Viagra) corrects $\Delta F508$ -CFTR location in nasal epithelial cells from patients with cystic fibrosis’ [77]. The publication states that the phosphodiesterase 5 inhibitor sildenafil increases $\Delta F508$ -CFTR trafficking. This means that the retention of the $\Delta F508$ -CFTR protein within the cell, which is typical for cystic fibrosis patients, can be corrected by administering sildenafil. The target within the publication, and therefore the target relevant for the workflow, is the phosphodiesterase 5. Multiple drugs are annotated for this target in DrugBank, one of them being vardenafil, which is why the association cystic fibrosis - phosphodiesterase 5 - vardenafil is one of the results of the workflow. Vardenafil was brought to my attention during the validation of the results with repoDB (described in detail in the section 2.1.6) because it is included in the repoDB dataset. A Phase II study (the ClinicalTrials.gov identifier is NCT01002534⁵⁰) called ‘Does a Nasal Instillation of Vardenafil Normalize the Nasal Potential Difference in Cystic Fibrosis Patients?’ seems to be on hold since 2012. This means that the association made by the workflow between cystic fibrosis and vardenafil is a valid one and has been considered sufficiently interesting to pursue clinical studies, even though they are currently suspended. Such valid associations are what this work was aiming for.

However, within the current dataset of results valid associations are still in the minority, because at this point in the workflow it is still impossible to determine whether the triangulation truly makes sense. This is because even though information about the mechanism of action of the drug on the target is present (e.g. inhibitor), information on the etiology (e.g. upregulation) of the disease is not. To determine the kind of mechanism that is needed, information about the etiology is vital. If for example a disease is due to an upregulation, an inhibitor is a useful drug, an activator however is not. There is no database which provides this information in a format that could be easily mined (for example in the form of annotations). Information on the etiology is only present in textual form, which

⁵⁰ <https://clinicaltrials.gov/ct2/show/NCT01002534?term=vardenafil&cond=Cystic+Fibrosis&rank=1>, accessed on 16.04.2018

is why I expanded the workflow and included an intermediate step, described in the next section.

2.3.4 Text mining

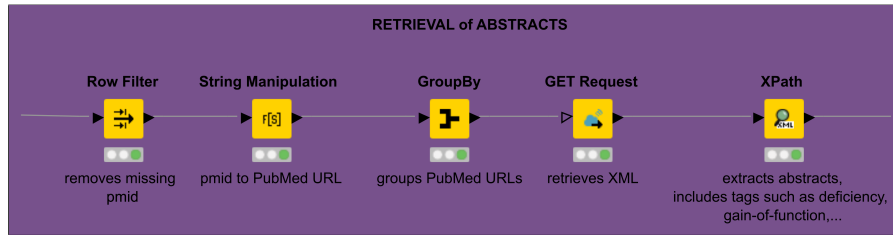


FIGURE 27: Nodes for the retrieval of publication abstracts based on PubMed identifiers

The last missing piece for the assessment of the suitability of the resulting drug-disease associations for repurposing was information on the cause of the disease. If the cause of the disease is an upregulation of a certain target, an inhibitor is needed. If the disease is caused by a downregulation of a certain target, an activator is needed. If however the disease is caused by a deficiency of a target, the results of the workflow are for the main part invalid because neither an activator nor an inhibitor are helpful in such a case, instead the therapy generally consists of an intravenous administration of the respective protein (which is also one of the reasons for the exorbitant costs of treatment for rare diseases, as described in section 1.1). To determine the cause of about 5,000 diseases in terms of up- and downregulation was however no small feat because currently there are no databases available that provide this information. An obvious approach is a literature search where in most cases a scan of the abstract already provides enough information to draw conclusions. Therefore I implemented an approach based on text mining with the aim to automate the assignment of the cause for all of the diseases to narrow down the workflow results to the truly meaningful ones. The text mining section within the workflow aims to emulate the approach of a manual curation to determine the etiology of the disease (or the type of association between a disease and a specific target).

Generally text mining approaches range from relatively simple methods (such as the use of dictionaries to identify words of interest and to classify accordingly) to more elaborate approaches (employing for example taggers and converting documents to vectors to be able to perform mathematical operations and to classify accordingly). During my work on the thesis I set up a dictionary-based as well as a vector-based approach, both of which require the abstracts to be retrieved as shown in Figure 27. This intermediate step of the workflow does in fact not require the third major step described in the previous section (2.3.3) to be executed because the retrieval here is done based on publication identifiers that were extracted in the second major step of the workflow (Extraction of genes (2.3.2)). It is explained here because this is the stage of the workflow at which the text mining results are needed.

The first node within the *Text mining* part of the workflow, the *Row Filter*, continues where the extraction of genes (described in subsection 2.3.2 and shown

in Figure 22 on page 47) ended. Here only rows with a valid PubMed identifier are taken into account.

The *String Manipulation* node then uses the expression `join("https://www.ebi.ac.uk/europepmc/webservices/rest/search?query=ext_id:$pmid$,&resulttype=core")` to create the syntax for the API call for the retrieval of abstracts. The 'resulttype' is set to 'core' so that the resulting XML will also include the abstract of the publication based on the given PubMed identifier ('pmid'). For one of the publications where cystic fibrosis is linked to the phosphodiesterase 5A the corresponding URL is `https://www.ebi.ac.uk/europepmc/webservices/rest/search?query=ext_id:15618584&resulttype=core`. The *XPath* node then uses appropriate expressions to extract for example the name of the publication and the abstract.

For a simple text mining approach additional information can already be extracted with the *XPath* nodes. With a predefined dictionary of terms the user is interested in, XPath expressions can be employed to search through the text in the XML and to retrieve a count if a term is present. Examples of XPath expressions that are used in this workflow are shown in Figure 28, which depicts a part of the configuration settings of the *XPath* node.

XPath summary		
Column name	XPath query	Type
gain-of-function	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'GAIN-OF-FUNCTION', 'gain-of-function'), 'gain-of-function')])</code>	String(SingleCell)
activating	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'ACTIVATING', 'activating'), 'activating')])</code>	String(SingleCell)
up-regulation	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'UP-REGULATION', 'up-regulation'), 'up-regulation')])</code>	String(SingleCell)
loss-of-function	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'LOSS-OF-FUNCTION', 'loss-of-function'), 'loss-of-function')])</code>	String(SingleCell)
down-regulation	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'DOWN-REGULATION', 'down-regulation'), 'down-regulation')])</code>	String(SingleCell)
deactivating	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'DEACTIVATING', 'deactivating'), 'deactivating')])</code>	String(SingleCell)
activation	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'ACTIVATION', 'activation'), 'activation')])</code>	String(SingleCell)
activated	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'ACTIVATED', 'activated'), 'activated')])</code>	String(SingleCell)
upregulation	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'UPREGULATION', 'upregulation'), 'upregulation')])</code>	String(SingleCell)
downregulation	<code>count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'DOWNREGULATION', 'downregulation'), 'downregulation')])</code>	String(SingleCell)

FIGURE 28: Examples of XPath expressions used to search for terms within an XML file

All of the expressions are based on the expression `count(/responseWrapper/resultList/result/abstractText[contains(translate(., 'UPPERCASE', 'lowercase'), 'lowercase')])`, which specifies that the *count* function shall be used (it retrieves the count if the specified term is indeed present in the XML file), the node within the tree on which the function shall be used, which in this case is the abstract entry (*abstractText*), and which terms to search for. The *translate* function is used here to ensure that terms in uppercase as well as in lowercase are identified. The resulting KNIME[®] table is shown in Figure 29. Cells which show a count for a term that was present in the abstract (here in all cases '1') are shaded.

Row ID	\$ pmid	\$ gain-of...	\$ activating	\$ up-reg...	\$ loss-of...	\$ down-r...	\$ deactiv...	\$ activation	\$ activated	\$ upregul...	\$ downre...	\$ down-r...	\$
Row84_1	11667971	0	0	0	0	0	0	0	0	0	0	0	0
Row85_1	11694004	0	0	0	0	0	0	1	0	0	0	0	0
Row86_1	11726399	0	0	0	0	0	0	0	0	0	0	0	0
Row87_1	11781704	0	0	0	0	0	0	0	0	0	0	0	0
Row88_1	11786964	0	0	0	0	0	0	1	1	0	0	0	0

FIGURE 29: KNIME[®] table resulting from a dictionary applied within an XPath node

Such a matrix can provide a basis for a classification, for example into diseases which arise due to an upregulation, downregulation or deficiency. However the method has two major drawbacks. The first one is that only terms which are present in the exact form as specified by the XPath expression will be identified (for example if the XPath expression specifies *upregulation*, the term *up-regulation* will not be identified), meaning that the approach requires a very extensive dic-

tionary. Secondly, the method fails whenever opposing terms are mentioned within the same abstract because the identification of terms does not consider the meaning the term was given, only if it is present or not. If a sentence were to be *The disease arises due to an upregulation of protein X and in some patients it has also been observed that protein Y is downregulated.*, this method results in a count of 1 for the terms *upregulation* and *downregulated*, resulting in a draw, whereas a classification would need a distinct difference.

The aforementioned sentence additionally illustrates a problem which is not specific to this method but to text mining itself, namely the correct extraction of directionality or relationships within the text. To the reader it is instantly clear that the term *upregulation* relates to *protein X* and the term *downregulated* to *protein Y*. Additionally the reader sees that the disease arises due to its relationship to *protein X*, not *protein Y*.

With a dictionary-based approach a classification of the abstracts can for example be performed with the nodes shown in Figure 30. These would be directly connected to the nodes shown in Figure 27. These nodes check how often the terms regarding upregulation, downregulation or deficiency occur, and whether one of them occurs more frequently than others. If they do, a category is assigned. If there is a draw, the abstract is not considered. Additionally, abstracts which refer to multiple genes and opposing types of associations are filtered out because within this approach, it is only possible to assign one category per abstract.

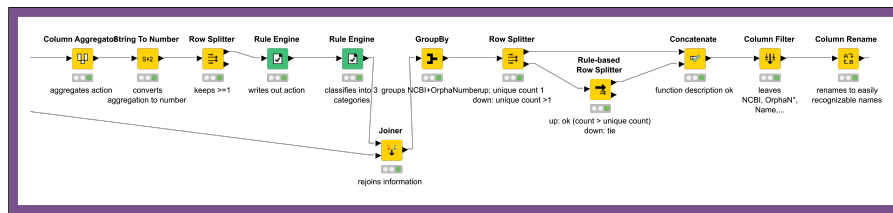


FIGURE 30: Nodes used for a classification of abstracts after a dictionary-based text mining

An example of a table resulting from this classification is shown in Figure 31. Three columns in this table originate from the classification, one for the original count, one where the occurrence of the terms is compared to each other, and one column for the final classification. The final classification (column ‘Cause of disease’) denotes the relationship between a disease (column ‘Orpha Name’) and a target (column ‘NCBI’). This classification is a prediction of the mechanism of action of a potential drug repurposing candidate.

Row ID	\$ OrphaName	\$ NCBI	\$ UMLS uri	\$ gain/loss of ...	\$ Cause of disease	\$ gain/loss of function (Count)	\$ Orpha Name
Row0	586	100289462	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(1)	gain-of-function	1	Cystic fibrosis
Row1	586	10043	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(1)	gain-of-function	1	Cystic fibrosis
Row2	586	100532736	http://linkedlifedata.com/resource/umls/id/C0010674	loss-of-function(1)	loss-of-function	1	Cystic fibrosis
Row3	586	10057	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(3)	gain-of-function	1	Cystic fibrosis
Row5	586	101929857	http://linkedlifedata.com/resource/umls/id/C0010674	loss-of-function(1)	loss-of-function	1	Cystic fibrosis
Row6	586	1022	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(1)	gain-of-function	1	Cystic fibrosis
Row7	586	10269	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(1)	gain-of-function	1	Cystic fibrosis
Row8	586	10365	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(2)	gain-of-function	2	Cystic fibrosis
Row10	586	10399	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(1)	gain-of-function	1	Cystic fibrosis
Row11	586	10595	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(1)	gain-of-function	1	Cystic fibrosis
Row12	586	10598	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(2)	gain-of-function	2	Cystic fibrosis
Row14	586	10724	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(1)	gain-of-function	1	Cystic fibrosis
Row16	586	10747	http://linkedlifedata.com/resource/umls/id/C0010674	deficiency(3)	deficiency	3	Cystic fibrosis
Row18	586	10993	http://linkedlifedata.com/resource/umls/id/C0010674	gain-of-function(3)	gain-of-function	1	Cystic fibrosis

FIGURE 31: Table resulting from a text mining approach based on a dictionary

The second approach to text mining that I implemented is shown in Figure 32. This approach is based on a vectorisation of the text and a subsequent classi-

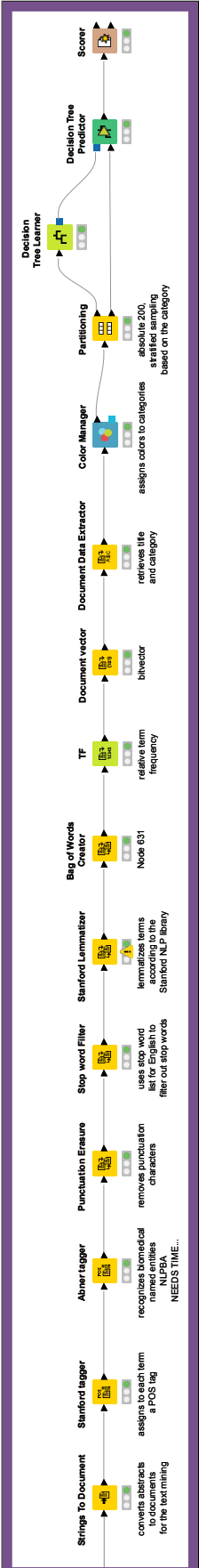


FIGURE 32: Nodes for a vector-based text mining approach

fication with an arbitrary classifier, such as random forest. For this approach the KNIME[®] extension for text mining has to be installed, as explained in section 2.2.3.

These nodes need publication abstracts as an input. Natively the abstracts are stored in the *string* format in KNIME[®], and the first step of the text mining extension is to convert this format to a *document* format (as seen in Figure 33) so that it can be processed by the text mining nodes.

\$ abstractText	\$ Title (First) (First)	\$ Orpha ...
BACKGROUND: Aquagenic wrinkling of the palms (AWP) is hyperwr...	Aquagenic wrinkling of the palms in cystic fibrosis and the cystic fi...	Cystic fibrosis
BACKGROUND: Cystic fibrosis (CF) is caused by compound heteroz...	The very low penetrance of cystic fibrosis for the R117H mutation:...	Cystic fibrosis
BACKGROUND: Cystic fibrosis (CF) results from mutations in the cy...	A novel natural product compound enhances cAMP-regulated chlor...	Cystic fibrosis
BACKGROUND: Cystic fibrosis is a monogenic recessive disorder fo...	Poly thymidine polymorphism and cystic fibrosis in a non-Caucasian...	Cystic fibrosis
BACKGROUND: Cystic fibrosis is caused by a defective gene encod...	Topical cystic fibrosis transmembrane conductance regulator gene ...	Cystic fibrosis
BACKGROUND: In airway epithelial cells, calcium mobilization can b...	Abnormal spatial diffusion of Ca2+ in F508del-CFTR airway epitheli...	Cystic fibrosis
BACKGROUND: Limited knowledge exists on phenotypes associate...	Non-classic cystic fibrosis associated with D1152H CFTR mutation.	Cystic fibrosis
BACKGROUND: Loss of cystic fibrosis transmembrane conductance...	Purinergic signaling underlies CFTR control of human airway epith...	Cystic fibrosis
BACKGROUND: Molecular diagnosis for cystic fibrosis (CF) is based...	Three novel CFTR polymorphic repeats improve segregation analy...	Cystic fibrosis
BACKGROUND: The defect in chloride and sodium transport in cysti...	Genetic analysis of Rwandan patients with cystic fibrosis-like symp...	Cystic fibrosis
OBJECTIVE: Primary hyperparathyroidism (pHPT)-related hypercal...	Pancreatitis risk in primary hyperparathyroidism: relation to mutati...	Cystic fibrosis
PURPOSE: Saguenay-Lac-Saint-Jean is a region located in the nort...	Distribution of CFTR mutations in Saguenay-Lac-Saint-Jean: prop...	Cystic fibrosis
1. Deletion of phenylalanine at position 508 (DeltaF508) of the cyst...	Identification of natural coumarin compounds that rescue defectiv...	Cystic fibrosis
A high number of mutations associated with cystic fibrosis have be...	Application of pyrosequencing to the identification of sequence va...	Cystic fibrosis
A new approach in the treatment of cystic fibrosis involves improvi...	Effect of VX-770 in persons with cystic fibrosis and the G551D-CFT...	Cystic fibrosis
A variety of mutation detection techniques, including restriction en...	Mutation characterization of CFTR gene in 206 Northern Irish CF f...	Cystic fibrosis

FIGURE 33: Table as seen after the conversion of publication abstracts from the *string* into the *document* format

The following nodes successively label and process the abstracts so that ideally in the end the core meaning of the text remains and can be encoded in the form of a vector. Firstly within the abstract each term is assigned a so-called ‘Part-of-Speech’ tag, such as noun or verb, by the *Stanford tagger*. The *Abner tagger* then assigns tags for biomedical entities, such as DNA. Next the node *Punctuation Erasure* removes punctuation marks such as comma. The *Stop word Filter* operates based on a list of stop words in English, such as ‘a’, ‘with’ or ‘not’. The *Stanford Lemmatizer* removes inflectional endings and returns the base or dictionary form of the word (known as lemma). The following example illustrates the lemmatization of a sentence:

the boy's cars are different colors
 $\Downarrow$
 the boy car be differ color⁵¹

Afterwards the terms from all of the documents are extracted by the *Bag of Words* node. The resulting table is shown in Figure 34. Each of the words in column ‘Term’ has a tag associated with it, assigned either by the *Stanford tagger*, in which case it is named ‘POS’ (specific examples are ‘NNS’ for *Noun, plural* or ‘VBD’ for *Verb, past tense*⁵²), or by the *Abner tagger*, in which case it says ‘ABNER’ and lists one of the five entities recognized by the Abner NLPBA corpus - protein, DNA, RNA, cell line and cell type⁵³.

51 <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>, accessed on 20.04.2018

52 For a complete list of entities consult *THE PENN TREE BANK: AN OVERVIEW* at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.8216&rep=rep1&type=pdf>, accessed on 20.04.2018.

53 Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**, 3191–3192 (2005).

T Term	Document
conclusion[NNS(POS)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
demonstrate[VBD(POS)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
suggest[VBZ(POS)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
decrease[NNS(POS)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
decrease[VBN(POS)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
CFTR protein[PROTEIN(ABNER)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
offer[VBN(POS)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
screen[VBG(POS)]	"aquagenic wrinkling palm cystic fibrosis cystic fibrosis carrier case-control stu...
low[JJ(POS)]	"low penetrance cystic fibrosis r117h mutation reappraisal genetic counselling ...
penetrance[NN(POS)]	"low penetrance cystic fibrosis r117h mutation reappraisal genetic counselling ...

FIGURE 34: Terms as extracted by the *Bag of Words* node

Now that all of the documents have been preprocessed, term frequencies can be extracted. The frequencies denote how often a term is found within a document and are necessary to vectorise all of the documents with the *Document vector* node. The resulting vector is shown in Figure 35. I included a *Document Data Extractor* afterwards to extract the title and the category of the abstract. This information is needed to train the classification models meant to predict the category of an abstract, as in the type of association between a disease and a target (e.g. deficiency).

Document	D CFTR[N...]	D Tgm[]	D tr[NIN(...]	D polymo...	D patient...	D CBAVD[...]	D populat...	D expres...
"CFTR Tgmtn polymorphism patient CBAVD population express low incidence cystic fibro...	2	1	1	2	8	1	2	2
"CFTR gene transfer human cystic fibrosis pancreatic duct cell use Sendai virus vector"	7	0	0	0	1	0	0	0
"CFTR mutation Algerian population"	2	0	0	0	3	0	4	0
"Complex allele -1023aa549T/G associate milder form cystic fibrosis allele p549tg"	0	0	0	0	8	0	0	0
"Hypoxia inducible factor-1 hif-1-mediated repression cystic fibrosis transmembrane con...	3	0	0	0	0	0	0	0
"N-terminal CFTR missense variant severely affect behavior cfr chloride channel"	3	0	0	0	1	0	0	0
"Nucleotide-binding domains cystic fibrosis transmembrane conductance regulator ABC t...	2	0	0	0	0	0	0	0
"Time-1 regulate recycling cystic fibrosis transmembrane conductance regulator"	7	0	0	0	0	0	0	0

FIGURE 35: Bitvector of terms within publication abstracts

The *Color Manager* was used during the testing of the approach to visualise the categories the dataset consists of. This was done to visually inspect whether the distribution of the categories is equal, so as not to bias the classification.

The final step of this approach is an arbitrary classifier (these nodes are not part of the KNIME[®] *Text Mining* extension). Shown in Figure 32 is a Random Forest classifier which consists of three nodes, one node for the partitioning of the data into a training set and a test set, one node for the generation of the tree and a final node for the prediction. The last node in Figure 32 is a scorer which I used to characterise the performance of the classification.

While testing this approach I also tried a Support Vector Machine classifier but had to discard it because the computational requirements (runtime and memory consumption) were very high.

Classifiers require a training set to classify correctly. To build a training set I used two different approaches.

My first approach was to manually select several diseases per class (e.g. gain of function, loss of function) which had a high number of associated publications. I then annotated these and used their publication abstracts as a training set. This method however did not yield satisfactory results.

My second approach was to use selected classes assigned to disease-gene relationships in the 'Rare diseases and cross-referencing' dataset and to use their publication abstracts as a training set. The difference to the above mentioned

approach is that here, a class is assigned to a disease-gene relationship, not to a disease as a whole as it is done within the first approach.

The classes included within the ‘Rare diseases and cross-referencing’ dataset are shown in Figure 36. The classes that I used for the generation of the training set are italicised.

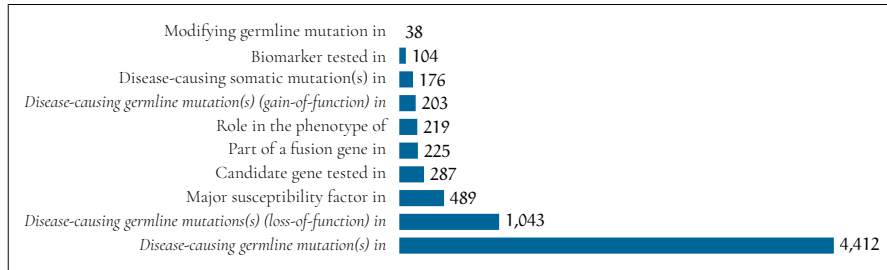


FIGURE 36: Disease-gene association types as seen within the ‘Rare diseases and cross-referencing’ dataset

One of the problems with this approach is that the dataset is unbalanced, meaning that there is a vast difference between the number of associations belonging to the three categories. The ratio unfortunately stays unbalanced even when publication abstracts are connected to the respective associations. However, there is a node in KNIME® called *Equal Size Sampling* which I used to balance the dataset. It extracts an equal number of abstracts from each category for the training set.

The second problem, which surfaced already during the first approach, is that publication abstracts are not always unique to a single disease-gene association. Sometimes one abstract refers to multiple genes or multiple diseases, meaning that one and the same abstract can be annotated with different classes, depending on the gene-disease association that is currently being examined. A single abstract with multiple classes is useless for a classifier, however the removal of such abstracts reduces the number of overall abstracts available for the creation of the training set.

Unfortunately, neither of the two approaches delivered entirely satisfying results. In regard to the second approach, the *Decision Tree Learner* node has an option of examining the tree that was created for the classification, and very often it is the case that the tree uses the name of a specific disease as a decision, which is meaningless for prediction across all diseases.

Without a reliable classification of publication abstracts disease-gene relationships have to be curated manually to determine the correct mechanism of action of a potential drug repurposing candidate. Given that the second text mining approach resulted in unreliable decision trees but results were needed to streamline the analysis of the drug repurposing results described in section 2.3.5, I chose to use the results from the first text mining approach.

2.3.5 Analysis of results

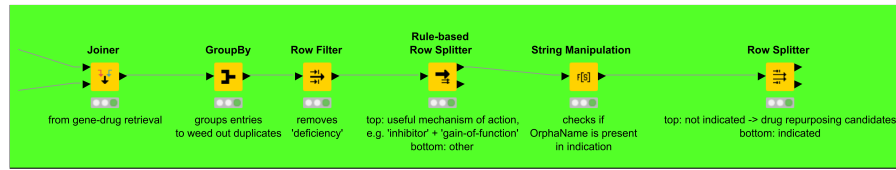


FIGURE 37: Nodes for the analysis of the workflow results and for the evaluation of drug repurposing candidates

The last part of the workflow, shown in Figure 37, joins the results from the text mining to the ones from the drug retrieval.

In a first step the results are grouped because there are duplicate entries left from the previous steps. Then, under the assumption that a deficiency cannot be cured with a small molecule drug, all the disease-gene associations that have been assigned the category ‘deficiency’ are removed from the dataset. The *Rule-based Row Splitter* ensures that only useful combinations of disease classes and mechanism of action of drugs are retained. The DrugBank XML (the mining of which was described in section 2.3.3) contains 54 mechanisms of action in total. However, not all of them are of interest to the workflow (such as ‘chaperone’). The interesting ones were used in expressions within the *Rule-based Row Splitter* (shown in Listing 6) to retain correct combinations only. An ‘antagonist’ for example is useful when the disease is due to a ‘gain-of-function’ of the respective protein, such a combination is therefore retained.

```

($action$ MATCHES "Antagonist|antagonist") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "activator") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "agonist") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "antibody") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "blocker") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "inactivator") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "inducer") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "inhibitor") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "negative modulator") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "neutralizer") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "partial agonist") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "partial antagonist") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "positive allosteric modulator") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "positive modulator") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "potentiator") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "reducer") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "stimulator") AND ($Cause of disease$ MATCHES "loss-of-function") => TRUE
($action$ MATCHES "suppressor") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE
($action$ MATCHES "weak inhibitor") AND ($Cause of disease$ MATCHES "gain-of-function") => TRUE

```

LISTING 6: Expressions used in the *Rule-based Row Splitter* to retain only correct combinations of disease-gene associations and the mechanisms of action of drugs

After this step, the *String Manipulation* node checks whether the name of the orphan disease is present in the indication for the drug that was provided by the DrugBank XML. The *Row Splitter* then splits the dataset accordingly. The bottom of the node contains entries where the name of the disease is present in the indication, meaning that the connection between the disease and the drug is already known and therefore not suitable for drug repurposing. The top contains entries of possible repurposing candidates as determined by the workflow.

RESULTS

This chapter outlines the results of the workflow as well as the results from the employed validation methods.

3.1 WORKFLOW RESULTS

The counts at positions within the workflow where datasets are introduced and analysed are shown at the end of this chapter in tabular format in Table 8. The colours correspond to the steps of the workflow as shown in Figure 18 on page 45.

The first column denotes the dataset (either downloaded or evaluated within KNIME®) for which the number of entries in the KNIME® table is shown in the column ENTRIES. The column ORIGIN denotes the origin of the dataset or the database it is related to.

For the first step of the triangulation the workflow starts out with the XML file ‘Rare diseases and cross-referencing’ which contains 9,557 diseases. Usually the number of rare diseases is estimated to be between 5,000 - 8,000, as described in the chapter 1. The high number of diseases in Orphanet is due to the fact that Orphanet differentiates between particular manifestations of diseases which might be classified as a single disease in the Unified Medical Language System (UMLS). For example in the case of ABeta amyloidosis, there is an Arctic, Dutch, Iowa and Italian type. All of them have different Orphanet identifiers, but only the Dutch type has a UMLS identifier. These cases, coupled with instances where no UMLS identifiers were provided, result in a total of 5,605 UMLS identifiers which are used further in the workflow. Given that one disease can also have multiple UMLS identifiers, the total number of unique diseases captured here is slightly lower at 4,838.

The second step of the triangulation - the retrieval of target information - is done in two parts. At first the retrieval of targets is done through Open PHACTS. This is done based on the 5,605 UMLS identifiers from the ‘Rare diseases and cross-referencing’ dataset, resulting in 113,636 unique disease-gene associations. These are annotated with the association type (an example of such an annotation is in Line 4 of Listing 4 in subsection 2.3.2). Multiple types for one disease-gene association are possible if multiple publications refer to the association. Association types and their counts in the full dataset (which has separate entries for each disease-gene association and its corresponding publication) are shown in Figure 38.

Secondly the ‘Rare diseases with their associated genes’ dataset from Orphadata is accessed, which additionally introduces 3,661 unique diseases to the workflow. For these diseases 7,046 gene associations are available. In total the number of unique disease-gene associations that result from the combination of the two

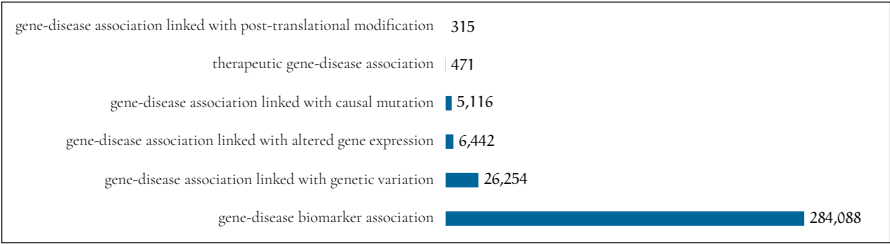


FIGURE 38: Types of disease-gene associations with respective counts

datasets is 117,851. In this dataset 11,170 genes are unique and they are connected to 6,314 unique diseases. For some of the diseases the mining results in up to several thousand associations. An overview of the distribution of the dataset is shown in Figure 39. The x-axis depicts the number of genes associated with a disease, and the y-axis depicts the number of diseases that are present in the dataset for a given number of genes. It can for example clearly be seen that there are a few diseases associated with a very high number of genes (such as one disease with 3,341 genes - the rare malignant breast tumour with the Orphanet identifier 180257, one disease with 2,187 genes - the hepatocellular carcinoma with the Orphanet identifier 88673, or a group of diseases with 2,065 genes - the rare diabetes mellitus group with the Orphanet identifier 101952). On the other hand a large number of diseases are associated with only a few genes (such as 3,237 diseases with a single gene - constituting 51.3% of the diseases, 683 diseases with two genes - constituting 10.8% of the diseases, or 308 with 3 genes - constituting 4.9% of the diseases). The densely populated area of the plot is shown in higher detail in Figure 40. The thresholds of a maximum of 20 diseases and a maximum of 300 genes were chosen here purely for reasons of visibility. Both plots clearly show a distribution where many diseases will have a few genes associated with them, as opposed to a few diseases with several hundreds or thousands of associated genes.

For the third step of the triangulation drugs are retrieved based on targets. Out of the 11,170 unique genes, information related to drugs is successfully retrieved through Open PHACTS for 10,990 of them. Drugs are available for 1,120 of these genes (more specifically for their respective protein products, known as targets), and for 770 of the genes 1,291 approved drugs are available. It is possible for a gene to have multiple approved drugs and for one drug to be associated with multiple genes, therefore the number of unique gene-drug associations captured in the dataset amounts to 4,332.

The triangulation is now therefore complete and results in 86,195 unique disease-drug associations. In this dataset there are 2,075 unique diseases and 1,281 unique drugs. This means that currently the workflow suggests 1,281 drugs as possible drug repurposing candidates for roughly 19% of the diseases the drug retrieval started out with.

To ensure that the proposed disease-drug combinations truly make sense (for example the disease-gene association is due to a gain of function and the drug is an inhibitor) text mining is introduced into the workflow and 137,072 publication abstracts are retrieved. Each of these abstracts is then assigned a cat-

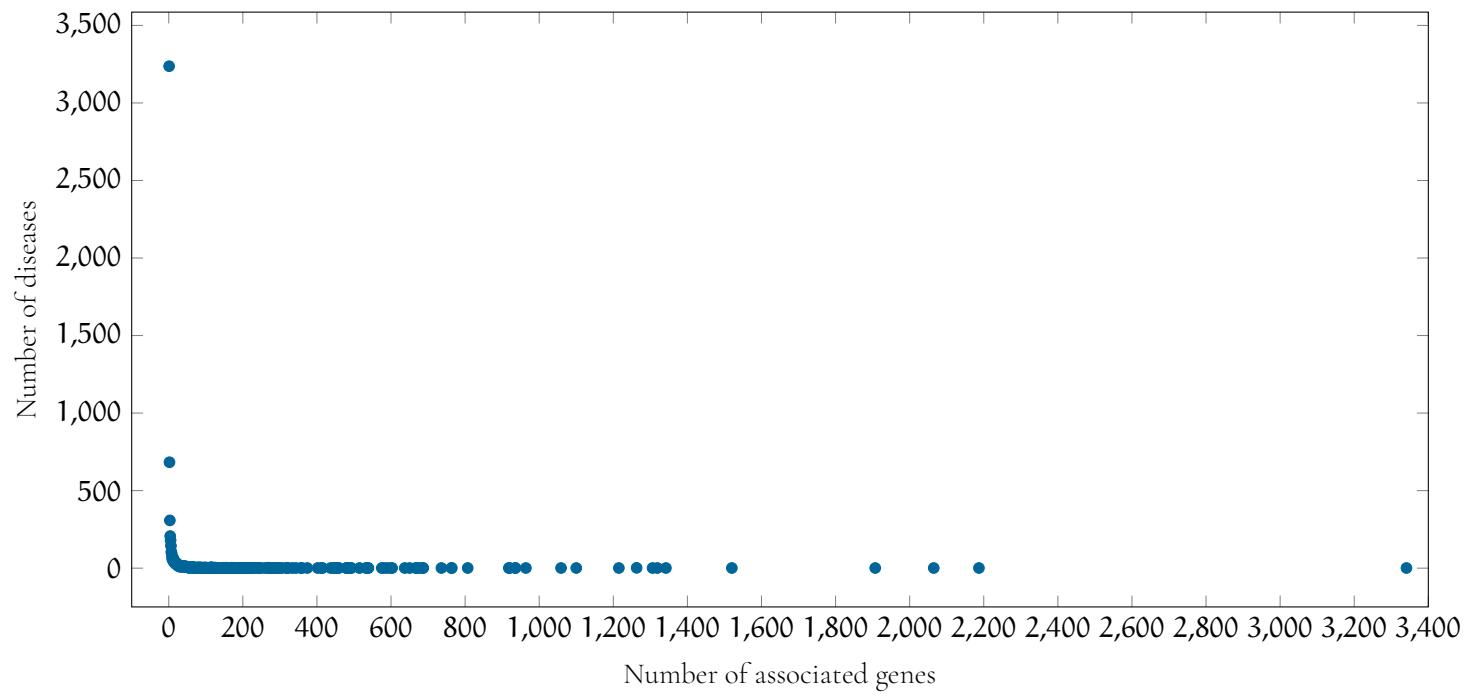


FIGURE 39: Distribution of the number of diseases per associated number of genes

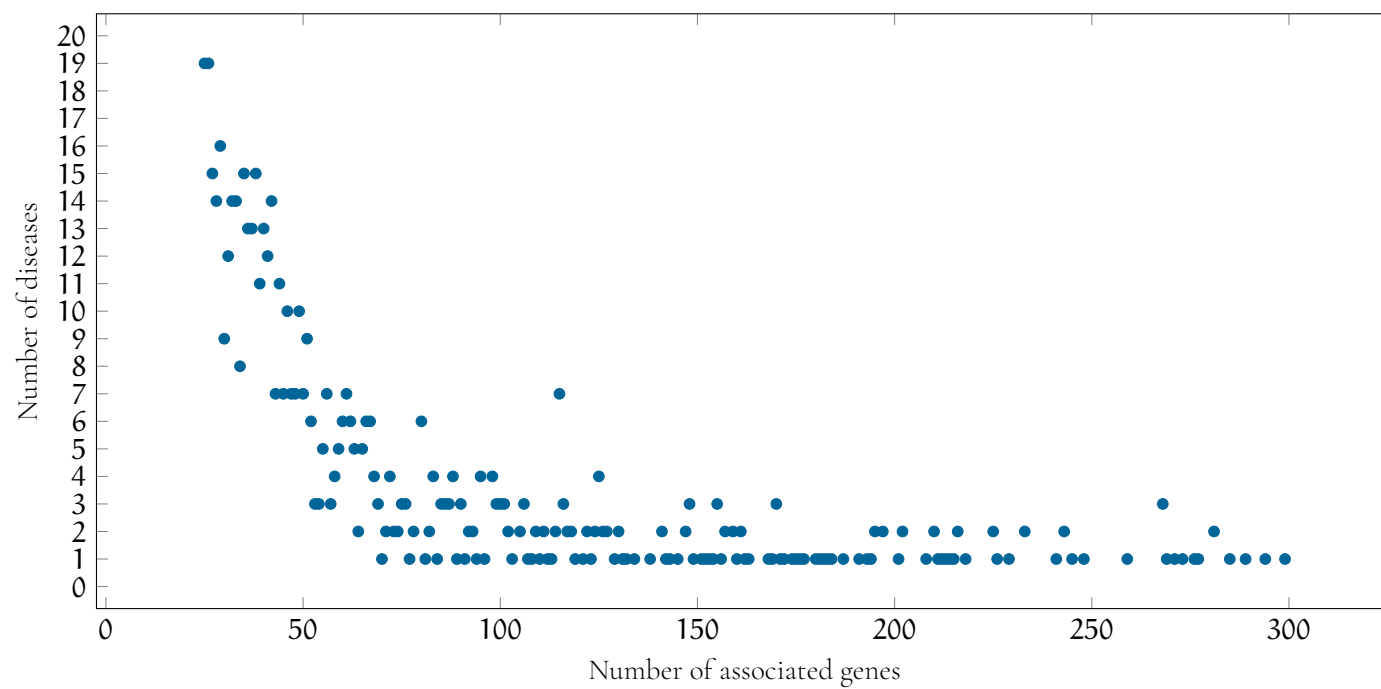


FIGURE 40: Distribution of the number of disease per associated number of genes for a maximum of 20 diseases and 300 genes



FIGURE 41: Distribution of unique disease-drug associations and their categories as assigned by text mining

egory ('gain-of-function', 'loss-of-function', 'deficiency') based on the approach described in subsection 2.3.4 and the categories are evaluated across the corresponding diseases and genes. This evaluation results in 53,078 abstracts with valid categories. The distribution of the categories is shown in Figure 41. The number of unique disease-drug associations in this dataset amounts to 43,182.

Assuming that small molecules are rather unsuitable to cure an etiology involving a deficiency, the dataset is then reduced to entries annotated with the categories 'gain-of-function' and 'loss-of-function', resulting in 37,260 unique associations between a disease and a drug.

The part of this dataset which has a useful combination of the etiology of the disease (based on the respective protein) and the mechanism of the drug, such as a 'gain-of-function' category for the etiology and a drug annotated as an 'inhibitor', amounts to 19,959 unique entries within the workflow.

The last step of the analysis checks whether the name of the disease is present in the indication description of the drug (to check whether the treatment is already known) and keeps only entries for which this is not the case, meaning that the use of the drug for the disease is currently not annotated in DrugBank. This leads to 19,924 unique disease-drug associations.

The number of unique disease that are captured here is 811 and the number of unique drugs is 1,003. This means that for some diseases (such as the rare malignant breast tumour mentioned above), the workflow suggests up to several hundred drug repurposing candidates.

TABLE 8: Results at different positions within the workflow

	DATASET	ORIGIN	ENTRIES	SECTION	DESCRIPTION
A	Rare diseases and cross-referencing	Orphadata	9,557	2.3.1	Dataset with Orphanet diseases and identifiers for other databases.
B	Valid UMLS identifiers	A	5,605	2.3.1	Identifiers within the <i>Rare diseases and cross-referencing dataset</i> that were suitable for a connection to Open PHACTS.
C	Unique diseases	B	4,838		Unique diseases contained within the dataset of valid UMLS identifiers.
D	Unique disease-gene associations	Open PHACTS (DisGeNET) <i>based on B</i>	113,636	2.3.2	Unique associations between diseases and genes that were retrieved through Open PHACTS based on valid UMLS identifiers.
E	Rare diseases with their associated genes	Orphadata	3,661	2.3.2	Disease entries in the <i>Rare diseases with their associated genes</i> dataset.
F	Unique disease-gene associations	Orphadata <i>based on E</i>	7,046		Unique associations between diseases and genes within the <i>Rare diseases with their associated genes</i> dataset.
G	Unique disease-gene associations	D + F	117,851		Unique associations between diseases and genes that resulted from the combination of the associations retrieved through Open PHACTS and from the <i>Rare diseases with their associated genes</i> dataset.
H	Unique genes	D + F	11,170		Unique genes resulting from the combination of the results retrieved through Open PHACTS and from the <i>Rare diseases with their associated genes</i> dataset.
I	Unique diseases	C + E	6,314		Unique diseases resulting from the combination of the diseases with valid UMLS identifiers retrieved from the <i>Rare diseases and cross-referencing dataset</i> and diseases from the <i>Rare diseases with their associated genes</i> dataset.
J	Genes with successful <i>Target</i> API calls	Open PHACTS <i>based on H</i>	10,990	2.3.3	Genes for which entries were present in Open PHACTS.

TABLE 8 Continued: Counts at different positions within the workflow

	DATASET	ORIGIN	ENTRIES	SECTION	DESCRIPTION
K	Genes with drugs	Open PHACTS (DrugBank) <i>within J</i>	1,120		Genes for which drugs were annotated in Open PHACTS.
L	Genes with approved drugs	K	770		Genes for which approved drugs were annotated in Open PHACTS.
M	Unique approved drugs	L	1,291		Unique drugs within the approved drugs retrieved for genes.
N	Unique gene-drug associations	$H + M$	4,332	2.3.3	Unique associations between genes and approved drugs, retrieved through Open PHACTS based on the genes retrieved through Open PHACTS and from the <i>Rare diseases with their associated genes</i> dataset.
O	Unique disease-drug associations <i>before text mining</i>	$I + M$	86,195		Unique associations between diseases and drugs which were retrieved based on the genes associated with the diseases, preceding the inclusion of publication abstracts.
P	Unique diseases <i>before text mining</i>	O	2,075		Unique diseases contained within the dataset of unique disease-drug associations before text mining.
Q	DRUG REPURPOSING CANDIDATES <i>before text mining</i>	O	1,281		Unique drugs contained within the dataset of unique disease-drug associations before text mining.
R	Unique publication abstracts	EMBL-EBI (Europe PMC)	137,072	2.3.4	Abstracts of publications retrieved from PubMed Central based on the identifiers of publications which were retrieved through Open PHACTS for disease-gene associations.
S	Abstracts with valid categories	<i>within R</i>	53,078	2.3.4	Associations between a disease, its gene and a related drug, with a category for the etiology of the disease ('gain-of-function', 'loss-of-function', 'deficiency') assigned by the text mining approach.

TABLE 8 Continued: Counts at different positions within the workflow

	DATASET	ORIGIN	ENTRIES	SECTION	DESCRIPTION
T	Unique disease-drug associations <i>after text mining</i>	$I + M$ considering S	43,182	2.3.5	Unique associations between diseases and drugs after the removal of conflicting categories.
U	Unique disease-drug associations <i>without the category 'deficiency'</i>	$T + \text{constraint}$	37,260	2.3.5	Unique associations between diseases and drugs without those with the category 'deficiency', meaning that the dataset now includes only the categories 'gain-of-function' and 'loss-of-function'.
V	Unique disease-drug associations <i>with useful combinations of a category and a drug's mechanism of action</i>	$U + \text{constraint}$	19,959	2.3.5	Entries for which the combination of the association between the disease and a gene and the mechanism of action of a drug are useful for potential drug repurposing (such as 'gain-of-function' and 'inhibitor').
W	Unique disease-drug associations <i>after text mining</i>	$V + \text{constraint}$	19,924	2.3.5	Unique associations between diseases and drugs without those mentioned in the description of the drug's indication provided by DrugBank.
X	Unique diseases <i>after text mining</i>	<i>within W</i>	811		Unique diseases at the end of the workflow.
Y	DRUG REPURPOSING CANDIDATES <i>after text mining</i>	W	1,003		Unique approved drugs at the end of the workflow, which can be considered as possible drug repurposing candidates.

3.2 VALIDATION

Two validation strategies were employed to validate the workflow results. An SV, a sensitivity-based validation only, and an SSV, a both sensitivity- and specificity-based validation. The validation strategies for *in silico* drug repurposing are described in detail in the section Validation strategies for computational drug repurposing (1.3).

Both validation methods yield matching results and suggest that roughly two out of three associations retrieved by the triangulation method are indeed viable for drug repurposing. However, the final validation datasets were specific subsets of the workflow results, possibly introducing a bias to the evaluations.

3.2.1 Validation with the Rare Disease Repurposing Database

For the SV (sensitivity-based validation) the Rare Disease Repurposing Database, previously offered by the FDA, was used. The database is described in the section Rare Disease Repurposing Database (2.1.5). The validation is based on the Excel sheet with products with an approval for a common disease indication. The database was discontinued in early 2017 and the validation was performed when the database was still live, meaning that the results reported here are based on an earlier version of the workflow.

The drugs in the Excel sheet were compared to the results from a version of the workflow where only the highest ranked genes per disease were taken into account, resulting in 205 associations proposed by the workflow between an orphan disease and a drug listed in the Excel sheet. All of these associations were then manually checked for their validity. Each of the identifiers was checked in the corresponding database, the associations originating from publications (disease-gene) were examined closely to determine their veracity (at times the associations are merely artefacts of the underlying mining method) as well as the etiology of the disease, and the associations to drugs were then checked for their usefulness (to ensure that the etiology of the disease matched the drug's mode of action). Lastly, for associations deemed to be valid, it was checked whether there were any publications discussing the given disease-drug association or even finished or ongoing clinical trials. The last step was done to determine to what extent the proposed association was already known or had been examined by the scientific community.

The results from this validation are shown in Figure 42. In 119 (58.0%) cases the associations were determined to be useful. These were associations which were present in scientific publications, associations with a potentially beneficial triangulation, associations that were or had been under investigation in clinical trials and one association with an approval by the FDA. In some cases the usefulness could not be determined and roughly 28% of the associations were determined not to be useful. All in all roughly 60% of the results were determined to be viable for drug repurposing.

3.2.2 Validation with repoDB

For the SSV (both sensitivity- and specificity-based validation) repoDB, introduced by Brown & Patel, was used. The database is described in the section repoDB (2.1.6). The database consists of both true positives (approved disease-drug combinations), and true negatives (failed disease-drug combinations) originating from DrugCentral¹ and the Aggregate Analysis of ClinicalTrials.gov Database².

A comparison of the results from the workflow and repoDB resulted in a subset of 331 associations which were present in both the workflow and repoDB. Of the 331 overlapping disease-drug associations 60.4% were approved, showing an enrichment comparable to the SV validation (58.0%). Roughly 40% of the overlapping associations failed in their attempt at repurposing during clinical trials, being either suspended, terminated or withdrawn due to various reasons. Interestingly, of the 102 failed associations annotated with descriptions, the most common reason for failure was a lack of enrolment of patients. Only a handful of trials failed due to safety or efficacy concerns. Detailed figures from the analysis are shown in Figure 43.

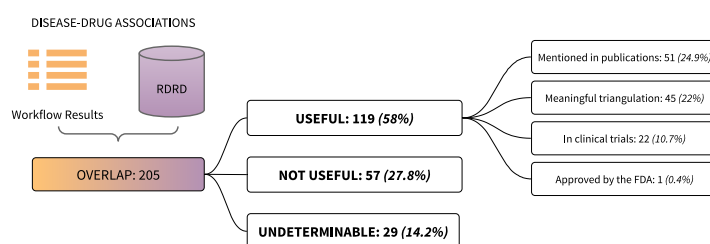


FIGURE 42: Results from the sensitivity-based validation with the Rare Disease Repurposing Database

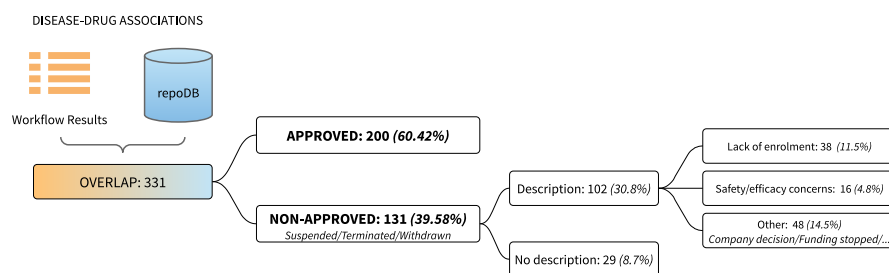


FIGURE 43: Results from the sensitivity- and specificity-based validation with the repoDB

¹ <http://drugcentral.org/>, accessed on 21.01.2018

² <https://aact.ctti-clinicaltrials.org/>, accessed on 21.01.2018

DISCUSSION

Results from the validation strategies for the workflow suggest that roughly two out of three proposed drug repurposing candidates are viable. This means that the aim of this work, which was to create a workflow capable of proposing possible drug repurposing candidates for orphan diseases, was achieved. The workflow not only connects orphan diseases to targets and drugs, it also retrieves valid drug repurposing candidates.

4.1 LIMITATIONS

Given that this work was done with state of the art databases relevant to orphan diseases and drugs and the results before the text mining (which reduces the dataset further, as described in chapter 3) cover roughly 2,000 orphan diseases, the statement by Muthyala in ‘Orphan/rare drug discovery through drug repositioning’ [26], that the current rate of drug repurposing activities raises the expectation that a substantial percentage of rare diseases, if not all 8,000, might be treatable with drugs in the current pharmacopoeia, seems doubtful. The limitation seems to be the number of available approved compounds suitable for the treatment of orphan diseases, which within the results preceding the text mining amounts to roughly 1,300 (the entire DrugBank contains 3,272 compounds which are annotated as ‘approved’¹). This number of approved compounds, seen in the context of 8,000 orphan diseases, indicates that new drugs will have to be developed to meet the need of the rare disease patient community.

The second limitation regarding drug repurposing for orphan diseases is the etiology of the disease. The limitation manifests in two ways. One is the fact that diseases where patients suffer from a protein deficiency are often treated with biologicals instead of small molecules, meaning that the treatment consists of the specific protein that is missing, necessitating dedicated drug discovery and development. By the estimation of the text mining approach, approximately every tenth² orphan disease arises due to a deficiency, meaning that every tenth orphan disease might need dedicated drug discovery and development. The second manifestation is that the etiology is vital for an automated assignment of drug repurposing candidates, but currently it can only be assigned manually with a high degree of certainty. Perhaps a superior text mining approach might lead to a reliable enrichment of drug repurposing candidates in future versions of this work.

In those cases where a drug repurposing candidate were to be found, patients might still have to settle for an off-label use, as seen within the SSV validation results (shown in Figure 43) which suggest that roughly every tenth drug repur-

¹ The number was obtained by mining the DrugBank 5.0 XML file.

² The number was derived from Table 41.

posing attempt for an orphan disease fails due to a lack of enrolment in clinical studies.

4.2 POSSIBILITIES FOR ADAPTATION

Within the workflow there are several positions at which it can be adapted to deliver targeted results. They are located within the gene retrieval segment of the workflow, the drug retrieval segment and the text mining segment.

4.2.1 *Gene retrieval segment*

Three possibilities for adaptation are for example located within the gene retrieval segment (described in subsection 2.3.2). The possibilities are to filter for specific disease classes, to choose a single target per disease and to filter based on the type of association between a disease and a gene.

4.2.1.1 *Specific disease classes*

One of the possibilities is to filter for specific disease classes. The results retrieved through Open PHACTS include disease classes indexed by Medical Subject Headings (MeSH), such as ‘Cardiovascular Diseases’, ‘Immune System Diseases’ or ‘Parasitic Diseases’. The dataset might therefore be filtered for a specific disease class, or specific disease classes (such as ‘Bacterial Infections and Mycoses’ and ‘Parasitic Diseases’) could be removed from the dataset. The latter could be done to increase the validity of the retrieved targets. The target mining performed by the workflow results in human targets only. In case of infectious diseases human targets can for example be mentioned in regard to symptoms or the severity of the disease (see for example the paper ‘Polymorphisms of TNF-enhancer and gene for Fc γ RIIa correlate with the severity of falciparum malaria in the ethnically diverse Indian population’ [79]), not as the primary target. This then means that the connections the workflow makes to drugs are based on the human targets, which in such cases are more suited to a treatment of the symptoms instead of the cause of the disease. Therefore a removal of disease classes related to infections might increase the focus of the dataset.

4.2.1.2 *Single target per disease*

The second possibility is to choose a single target per disease, as opposed to all targets that were retrieved. Contrary to the general teachings of pharmacology, where the cause of a given disease is the respective gene and its protein product (e.g. an upregulation, deficiency, loss-of-function mutation), the data mining as executed with this workflow can at times retrieve up to several thousand associations for a single disease. Orphanet is curated manually, meaning that even though there may be multiple gene entries per disease, the number is usually quite limited. Open PHACTS however uses DisGeNET as the database for genes associated with diseases, and DisGeNET is based on curated sources as well as text mining. Text mining as a source means that whenever a disease and a gene

are mentioned in a publication, an association may be present in the database due to the nature of the method, even if the meaning within the publication was different and the gene may in fact not be of any therapeutic value for this particular disease. An example of such a case is the association between cystic fibrosis and dynactin subunit 4. This gene is scored third for cystic fibrosis (based on a current search in DisGeNET³), directly after the cystic fibrosis transmembrane conductance regulator and the transforming growth factor beta. Two publications are referenced for this association^{4,5}, both of which are devoted to phenotypic variants of interacting modifiers which were discovered in cystic fibrosis patients with chronic *Pseudomonas aeruginosa* infections. Dynactin subunit 4 therefore may be of therapeutic value for the treatment of chronic *P. aeruginosa* infections in cystic fibrosis patients, but not for the treatment of cystic fibrosis itself.

There are several possibilities to handle such cases. The scores DisGeNET provides for the associations (based for example on the type of journal the article was published in) could be used to filter the dataset, if a meaningful threshold can be determined. Unfortunately, the scores are not accessible through Open PHACTS and would require an incorporation of the native DisGeNET dataset into the workflow. This however also means that the entire workflow could be remodelled to use the native datasets only instead of Open PHACTS, given that also additional DrugBank data had to be added manually, and Open PHACTS is used within the workflow solely for the purpose of accessing DisGeNET and DrugBank data. I did such a manual dataset addition for DrugBank only because it was essential for the assignment of valid disease-drug categories for the end results of the workflow. Due to timely constraints of the thesis I did not remodel the entire workflow to include native datasets only.

Another possibility is the use of the highest ranked gene association based on the number of publications supporting it. The idea here is to take only the genes into account that have the highest number of publications. This however still poses a problem whenever two or more genes are associated with the same number of publications, but more importantly, the number itself does not constitute an automatic validation of a disease-gene association, and in choosing this approach the associations that have only recently been discovered and simply did not have many publications would be lost, as well as associations for those diseases where multiple targets are evidenced. However this approach might also increase the focus of the dataset and reduce the number of proposed repurposing candidates so that a manual curation of the results becomes feasible. Currently a manual curation is impossible for the entire dataset due to the high number (roughly 20,000) of disease-drug associations the workflow proposes for drug repurposing.

-
- 3 <http://www.disgenet.org/web/DisGeNET/menu/browser/tab8b?5&pview=default&pf=http://www.disgenet.org/web/DisGeNET%3Fdata/diseases::C0010674::de&pf=/data/sources::ALL::de>, accessed on 02.11.2018
 - 4 Emond, M. J. *et al.* Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nature genetics* **44**, 886 (2012).
 - 5 Emond, M. J. *et al.* Exome sequencing of phenotypic extremes identifies CAV2 and TMC6 as interacting modifiers of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *PLoS genetics* **11**, e1005273 (2015).

4.2.1.3 *Type of association*

The third possibility is to filter based on the type of association between a disease and a gene, for example by considering only associations annotated with ‘therapeutic gene-disease association’. Such a label is in subsection 2.3.2 in Listing 4 line 4. In this Listing the association is called ‘gene-disease biomarker association’. There are some associations within the dataset that are marked with ‘therapeutic gene-disease association’ (see also Figure 38), however they constitute only around 0.1% of the dataset and unfortunately they are still not entirely reliable (an association where the tumour necrosis factor is upregulated as a symptom might for example be labelled therapeutic, even though it is not the original cause of the disease). In light of these drawbacks I decided not to use any of the above mentioned constraints but to use the entire table for further processing.

4.2.2 *Drug retrieval segment*

Another position at which the workflow can be adapted is located within the drug retrieval segment (described in subsection 2.3.3). As this workflow was meant to suggest solely approved compounds as drug repurposing candidates, a filtering was implemented to discard other types of drugs, such as investigational, experimental or withdrawn drugs. If needed, the focus of the dataset could be changed here, so as to for example suggest withdrawn drugs, to investigate whether these could be rescued by a repurposing for an orphan disease.

4.2.3 *Text mining segment*

Finally the text mining segment (described in subsection 2.3.4) in its entirety harbours many possibilities of reducing or widening the dataset according to the aim. The results will first and foremost depend on the chosen text mining method. In addition to this, even with a simple dictionary-based method the results will greatly differ whenever the dictionary is changed and words are adapted, added or removed.

The work described here is a generalised approach for *in silico* drug repurposing for orphan diseases that successfully provided a dataset of possible candidates and several options for a targeted search. The work has however also brought several limitations to light which hamper efficient *in silico* drug repurposing (such as the lack of information regarding targets which are crucial for a disease and the lack of information regarding the etiology of a disease), as well as drug repurposing for orphan diseases in general (such as the lack of enrolment in clinical studies). I therefore conclude that drug repurposing for orphan diseases is by no means a trivial task, and requires combined efforts of the scientific as well as the patient community.

Part III

APPENDIX

The Appendix includes additional material that arose from this work, such as submissions for poster presentations and poster prizes at conferences and the corresponding posters.

POSTERS

A.1 LINKING LIFE SCIENCE DATA: DESIGN TO IMPLEMENTATION, AND BEYOND

This poster was presented at the conference *Linking Life Science Data: Design to Implementation, and Beyond*¹, which was held from 18th–19th February 2016 in Vienna as a closing conference for the Open PHACTS project. There were flash poster presentations of 1min each followed by a poster session.

The abstract and the poster were as follows.

¹ The Open PHACTS Discovery Platform. *Linking Life Science Data: Design to Implementation, and Beyond* Open PHACTS project closing conference. <<http://www.openphactsfoundation.org/linking-life-science-data/>> (2016).

Analysis of orphan diseases with a KNIME workflow using Open PHACTS, with the potential of drug repurposing

Jana Gurinova, Daniela Digles, Gerhard Ecker

University of Vienna, Department of Pharmaceutical Chemistry, Althanstraße 14, 1090 Vienna, Austria

Worldwide an estimated number of 400 million people are affected by orphan diseases [1]. An orphan disease is defined as affecting less than 1 in 2000 citizens [2]. Such a low prevalence coupled with the sheer number of orphan diseases, estimated to be about 5000-8000, is the main reason for the small number of marketing approvals, amounting to treatments for roughly 200 conditions in the US and only about 45 in the European Union [1]. Drug repurposing therefore embodies an attractive option of reaching many patients with treatments that have already been deemed safe. This work aims at providing an overview of linked targets and drugs for orphan diseases, with a special focus on involved transporters.

The biggest European platform for orphan diseases is Orphanet [3], with comprehensive information for patients as well as for professionals. Orphanet also provides identifiers for the listed diseases as well as genes, which are perfectly suitable for data integration across different databases and thus also data enrichment with the aim of understanding and visualizing the role of diverse protein classes such as ion channels or transporters in orphan diseases. The Open PHACTS Discovery Platform has proven invaluable for exactly this kind of data integration given that the included databases such as UniProt, DisGeNET and DrugBank brought together enable the user to analyse the data based on various criteria as well as to draw previously unseen conclusions.

The workflow itself was established by using the KNIME Analytics Platform and starts with extracting data from Orphanet such as the name of the disease and its UMLS identifier. This key component is then connected to the Open PHACTS Discovery Platform through the use of Open PHACTS nodes available from <https://github.com/openphacts/OPS-Knime>, and yields further information such as which targets are involved in the disease plus their respective UniProt identifiers, and ultimately the drugs for these targets and their stages of development including the ones that are already approved.

As of now, approximately a third of the 9000 disorders including their sub-types listed in Orphanet are equipped with diverse identifiers, 2901 of which are UMLS identifiers used in the workflow for further analysis. Additionally, previously unknown links of diseases to drugs may be investigated for their potential regarding drug repurposing.

Acknowledgements

The work has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. [115191], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution of EFPIA companies.

References

- [1] Kaplan W et al. Priority Medicines for Europe and the World Update 2013, *World Health Organization*, Geneva, 2013
- [2] EURORDIS. Rare Diseases: understanding this Public Health Priority. *EURORDIS*, Paris, 2005
- [3] Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at <http://www.orpha.net> (accessed December 2015)

Analysis of orphan diseases with a KNIME workflow using Open PHACTS, with the potential of drug repurposing

Jana Gurinova, Daniela Digles, Gerhard F. Ecker

University of Vienna, Department of Pharmaceutical Chemistry, Althanstraße 14, 1090, Vienna

Introduction

Worldwide around 400 million people are affected by orphan diseases [1], orphan meaning affecting less than 1 in 2000 citizens [2]. The low prevalence coupled with the sheer number of orphan diseases (about 5000-8000) is the main reason there are so few marketing approvals, amounting to treatments for roughly 200 conditions in the US and only about 45 in the European Union [1]. Drug repurposing therefore embodies an attractive option of reaching many patients with treatments that have already been deemed safe.

Aim

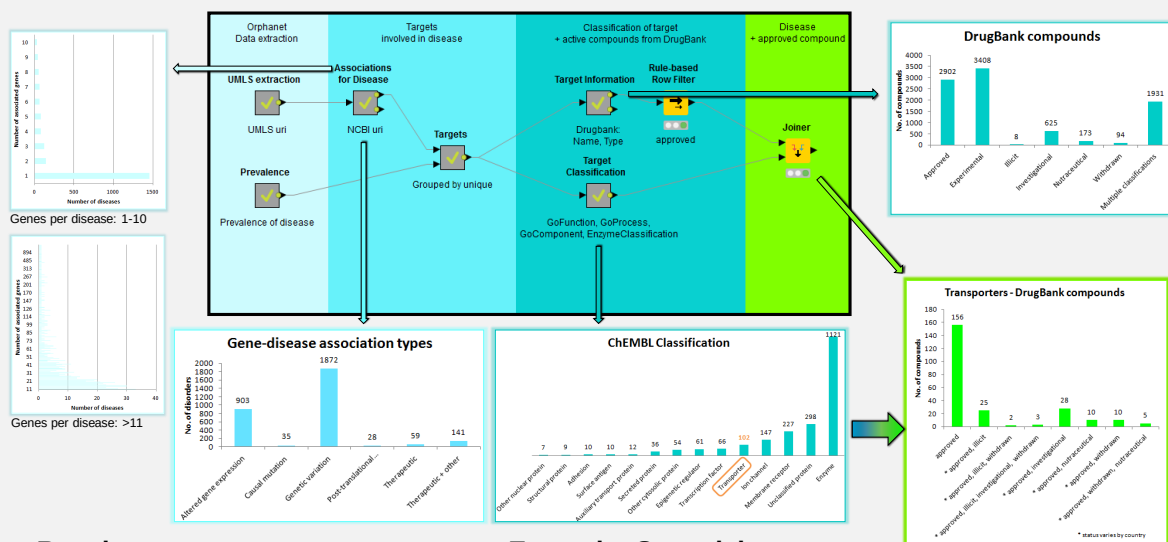
The workflow aims at providing an overview of linked targets and drugs for orphan diseases, with a special focus on involved transporters.

Methods

The workflow was created with the KNIME Analytics Platform software. For data retrieval from the Open PHACTS Discovery Platform nodes developed by Ronald Siebes (VU University Amsterdam), which are available from <https://github.com/openphacts/OPS-Knime>, were used.

The workflow

The biggest European platform for orphan diseases is Orphanet [3], and as of now, 2901 of the 9000 listed disorders including their sub-types are equipped with a UMLS identifier which was used in the workflow for retrieval of disease-related data from DisGeNET [4].



Results

In ChEMBL 102 of the targets are classified as transporters, whereas the search for transporters in the GO Molecular Function properties yields 311 transport related targets. Of the latter four are described as having a therapeutic gene-disease association and of these four, synovial sarcoma is the single one that was linked to approved compounds.

Conclusions

The workflow is capable of adapting to the user's needs. Options include filtering by highest cited or multiple genes for a given disease, the gene-disease association (e.g. therapeutic), or the compound type of interest (e.g. approved, investigational). Manual investigation of the resulting compounds and their pharmacological activity (agonist/antagonist) related to the disorder of interest, may result in possible drug repurposing candidates.

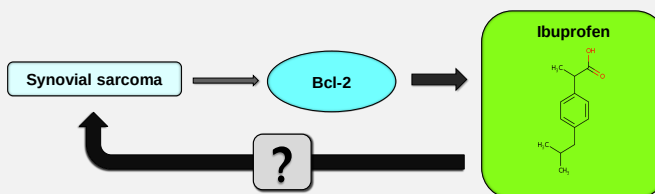
Acknowledgement:

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution of EFPIA companies.



Example: Synovial sarcoma

Given that transporters are of special interest to the research group, the results from the workflow (approved compounds only) were cross-examined with an established list of transport related proteins and yielded among other compounds Ibuprofen. Upon closer examination it became apparent that in synovial sarcoma the antiapoptotic factor Bcl-2 is frequently overexpressed, leading to a lower incidence of apoptosis. Ibuprofen interestingly seems to have the ability to downregulate the mRNA levels of Bcl-2, which would lead to enhanced apoptosis and could therefore have a therapeutic effect.



References:

- [1] Kaplan W et al. Priority Medicines for Europe and the World Update 2013, World Health Organization, Geneva, 2013
- [2] EURORDIS. Rare Diseases: understanding this Public Health Priority. EURORDIS, Paris, 2005
- [3] Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at <http://www.orpha.net>
- [4] Janet Piñero, Núria Queralt-Rosinach, Alex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, Laura I Furlong, DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (2015) Vol. 2015: article ID bav028; doi:10.1093/database/bav028

FIGURE 45: Poster for the Open PHACTS closing conference

A.2 THE ISMC AND THE YMCS

The second poster was presented at the *International Symposium on Medicinal Chemistry*² of the *European Federation for Medicinal Chemistry*, which was held from 28th August to 1st September 2016 in Manchester. The *International Symposium on Medicinal Chemistry* in Manchester was directly followed by the *Young Medicinal Chemist Symposium*³, taking place from 1–2nd September, also organized by the *European Federation for Medicinal Chemistry*. At both conferences this poster was chosen for a flash poster presentation of 5min each.

The abstract and the poster were as follows.

2 European Federation for Medicinal Chemistry. *International Symposium on Medicinal Chemistry* <http://www.ldorganisation.com/v2/produits.php?cle_menus=1238915829> (2016).

3 European Federation for Medicinal Chemistry. *Young Medicinal Chemist Symposium* <http://www.ldorganisation.com/produits.php?langue=english&cle_menus=1238916237> (2016).

TEMP309

TARGETS IN ORPHAN DISEASES AND THEIR REPURPOSING CANDIDATES - ANALYSIS WITH A KNIME WORKFLOW USING OPEN PHACTS

Jana Gurinova, Daniela Digles, Gerhard F. Ecker

University of Vienna, Department of Pharmaceutical Chemistry, Althanstraße 14, 1090 Vienna, Austria

Worldwide an estimated number of 400 million people are affected by orphan diseases¹. An orphan disease is defined as affecting less than 1 in 2000 citizens². Such a low prevalence coupled with the sheer number of orphan diseases, estimated to be about 5000-8000, is the main reason for the small number of marketing approvals, amounting to treatments for roughly 200 conditions in the US and only about 45 in the European Union¹. Drug repurposing therefore may prove to be the future of drug discovery for orphan diseases because it is an attractive option of reaching many patients with treatments that have already been deemed safe. This work aims at providing an overview of targets linked to orphan diseases as well as relevant compounds, consisting of possible repurposing candidates for these targets as well as experimental compounds as a starting point for drug discovery.

The biggest European platform for orphan diseases is Orphanet³, with comprehensive information for patients as well as for healthcare professionals, providing also identifiers (UMLS, MeSH, OMIM etc.) for 5345 of its 9235 listed diseases, as well as for associated genes. These identifiers are perfectly suitable for data integration across different databases and thus also for data enrichment with the aim of understanding and visualizing the role of diverse protein classes such as ion channels or transporters in orphan diseases. The Open PHACTS Discovery Platform contains the linked datasets for this kind of data integration, given that the included databases (such as DisGeNET, UniProt, DrugBank, SureChEMBL and ChEMBL) enable the user to go from the disease to its targets and ultimately to approved and experimental compounds connected to these targets. The approved compounds may include candidates for drug repositioning, whereas the experimental compounds may provide a foundation for drug discovery.

The workflow was established by using the KNIME Analytics Platform and starts with extracting data related to orphan diseases from different sources. Primarily it extracts identifiers provided by Orphanet and links them to respective targets through the use of DisGeNET and UniProt, additionally target references provided by Orphanet are also linked to the Open PHACTS Discovery Platform for further data integration, resulting in a combined dataset of specific targets for orphan diseases. The workflow then links this dataset to DrugBank, SureChEMBL and ChEMBL, with the results consisting of approved drugs as well as experimental compounds for the targets involved in orphan diseases.

It is remarkable that solely by using the identifiers from Orphanet the workflow already results in 1269 approved compounds that may be examined for suitable drug repurposing candidates.

Acknowledgements

The work has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. [115191], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution of EFPIA companies. We further wish to acknowledge the Open PHACTS Foundation, the charitable organisation responsible for the Open PHACTS Discovery Platform, without which this work would not have been possible.

References

- 1) Kaplan W et al., Priority Medicines for Europe and the World Update 2013, World Health Organization, Geneva, 2013
- 2) EURORDIS, Rare Diseases: understanding this Public Health Priority, EURORDIS, Paris, 2005
- 3) Orphanet: an online rare disease and orphan drug data base. © INSERM 1997. Available at <http://www.orpha.net> (accessed December 2015)

TARGETS IN ORPHAN DISEASES AND THEIR REPURPOSING CANDIDATES - ANALYSIS WITH A KNIME WORKFLOW USING OPEN PHACTS

JANA GURINOVA, DANIELA DIGLES, GERHARD F. ECKER

University of Vienna, Department of Pharmaceutical Chemistry, Althanstraße 14, 1090, Vienna

Introduction

Worldwide around 400 million people are affected by orphan diseases [1], the term 'orphan' meaning affecting less than 1 in 2000 citizens [2]. The low prevalence coupled with the sheer number of orphan diseases (about 5000-8000) is the main reason why there are so few market approvals, amounting to treatments for roughly 200 conditions in the US and only about 45 in the European Union [1]. Therefore, drug repurposing embodies an attractive option of reaching many patients with treatments that have already been deemed safe.

Aim

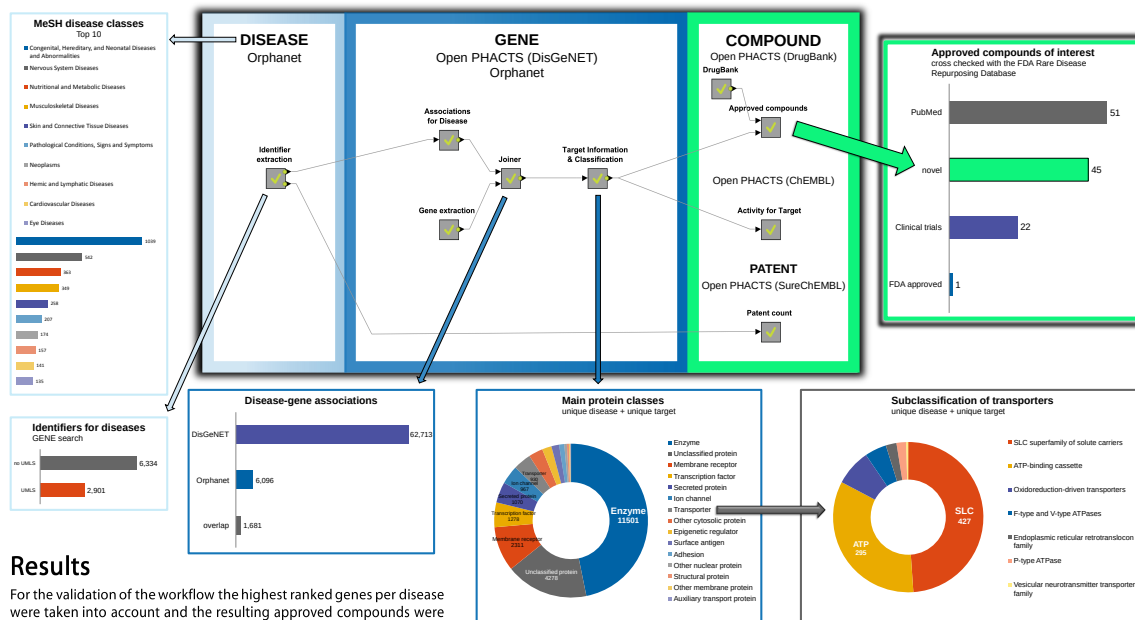
The workflow aims at providing an overview of linked targets and drugs for orphan diseases. Additionally it aims to serve as a starting point for drug repurposing.

Methods

The workflow was created with the KNIME® Analytics Platform software. For data retrieval from the Open PHACTS Discovery Platform nodes developed by Ronald Siebes (VU University Amsterdam), which are available from <https://github.com/openphacts/OPS-Knime>, were used. Additional data was retrieved in XML format from Orphadata [3] as well as from DrugBank [4].

The workflow

The biggest European platform for orphan diseases is Orphanet [5]. As of now, 2901 of the 9000 listed disorders including their sub-types are equipped with a UMLS Identifier which is used in the workflow for the retrieval of disease-related data from DisGeNET [6] to enrich the targets provided by Orphanet. Subsequently, information about the target and the associated compounds is retrieved and filtered for novel associations. Associations are taken into account if the compound is approved and its relationship to the function of the gene in the disease is beneficial (e.g. an approved inhibitor in case of a gain-of-function).



Results

For the validation of the workflow the highest ranked genes per disease were taken into account and the resulting approved compounds were cross checked with the FDA Rare Disease Repurposing Database. The obtained 205 disease-gene-compound associations were then manually examined for their viability, resulting in 119 associations of interest. 51 of the associations of interest are mentioned in PubMed, in 22 additional cases clinical trials are referenced and one combination is approved by the FDA. 45 associations were identified as novel and viable, the mentioned case study of ligneous conjunctivitis being one of them.

Conclusions

The workflow results in potential candidates for drug repurposing in addition to correctly retrieving associations already used therapeutically. There are multiple possibilities to adapt to the user's needs, such as filtering by highest cited or multiple genes for a given disease, the compound type of interest (e.g. approved, investigational) as well as filtering by patent count. Manual investigation of the resulting compounds may result in possible drug repurposing candidates.

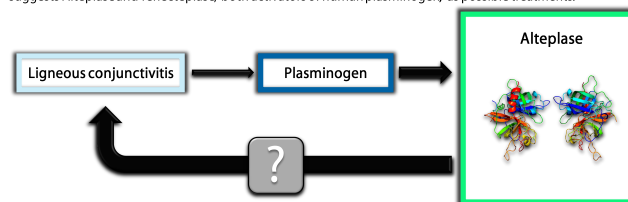
Acknowledgement:

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution of EFPIA companies.



Case study: ligneous conjunctivitis

Ligneous conjunctivitis is a rare form of chronic conjunctivitis characterized by the recurrent formation of pseudo membranous lesions most commonly on the eyelid surfaces. It is most frequently reported as a clinical manifestation of severe homozygous or compound-heterozygous hypoplasminogenemia and therefore, the treatment consists of human plasminogen. Because of the connection to human plasminogen, the workflow suggests Alteplase and Tenecteplase, both activators of human plasminogen, as possible treatments.



References:

- [1] Kaplan W et al. Priority Medicines for Europe and the World Update 2013, World Health Organization, Geneva, 2013
- [2] EURORDIS. Rare Diseases: understanding this Public Health Priority. EURORDIS, Paris, 2005
- [3] Orphadata: Free access data from Orphanet. © INSERM 1997. Available on <http://www.orphadata.org>.
- [4] Wishart DS et al. DrugBank. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72.
- [5] Orphanet: an online database of rare diseases and orphan drugs. Copyright INSERM 1997. Available at <http://www.orpha.net>
- [6] Piñero J et al. DisGeNET Database (2015) Vol. 2015: article ID bav028; doi:10.1093/database/bav028

FIGURE 47: Poster for the ISMC and YMCs

A.3 255TH ACS NATIONAL MEETING & EXPOSITION

This poster was presented at the 255th *American Chemical Society National Meeting & Exposition*⁴, which was held from 18th–22nd March 2018 in New Orleans. The poster was awarded the ACS Division of Chemical Information (CINF) Scholarship for Scientific Excellence. The abstract which was submitted for the conference, the abstract which was submitted to apply for the scholarship and the poster which was presented in the Sci-Mix session were as follows.

ABSTRACT SYMPOSIUM NAME: Sci-Mix

ABSTRACT SYMPOSIUM PROGRAM AREA NAME: CINF

CONTROL ID: 2869475

PRESENTATION TYPE: Poster Preferred : Consider for Sci-Mix

TITLE: Triangulation of repurposing candidates for orphan diseases

AUTHORS (FIRST NAME, LAST NAME): Jana Gurinova¹, Daniela Digles¹, Gerhard F. Ecker¹

INSTITUTIONS (ALL):

1. Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria.

ABSTRACT BODY:

Abstract: Worldwide approximately 400 million people are affected by orphan diseases. Orphan diseases are defined by their low prevalence and therefore a low number of patients. In the US a disease is considered to be rare when it affects less than 200000 patients, which reflects a prevalence of around 7 in 10000 citizens. Treatments for these conditions - of which there are around 5000-8000 - are however few and far between, amounting to around 500 drugs with a marketing approval for an orphan disease in the US. The need for new therapies can therefore be clearly seen. Coupled with the fact that an orphan drug can generate up to 818 million dollars in revenue over its competition-free lifespan, a trend can be seen where pharma's focus is shifting towards rare disease treatments.

A second recent trend is the repurposing of approved compounds. Especially the reduction of drug development costs by repurposing safe and approved compounds is a very beneficial aspect for the rare disease community, whose biggest problem is the low return on drug development costs because of the low number of patients.

To bring all these aspects together and facilitate a systematic search for repurposing candidates for orphan diseases, we developed a workflow which combines several databases and triangulates between the disease, its target and approved drugs for this target. This way we are able to find previously unnoticed associations between orphan diseases and approved drugs.

We start out with information on orphan diseases from Orphanet (a platform for orphan diseases), connect the diseases to their respective targets in DisGeNET (which we access through the Open PHACTS Discovery Platform), and connect the targets to approved drugs through DrugBank (a database with drug data and drug target information). The results of this triangulation are associations between diseases and approved drugs, comprised of previously known and more importantly also new associations. Currently we are working on streamlining the analysis of these associations so as to minimize the manual curation needed to arrive at viable repurposing candidates.

For some of the roughly 8000 orphan diseases, drug repurposing is the only hope of gaining access to approved therapeutics. The abovementioned method is a systematic approach that creates value for the patient community as well as for the pharma industry.

(No Image Selected)

FIGURE 48: Abstract for the 255th ACS National Meeting & Exposition

⁴ American Chemical Society. 255th ACS National Meeting & Exposition Accessed on 23.10.2017. <<https://global.acs.org/events/255th-acn-national-meeting-exposition/>>.

JANA GURINOVA
 University of Vienna, Austria
 jana.gurinova@univie.ac.at
 ID: CINF 66 (Control ID 2869475)

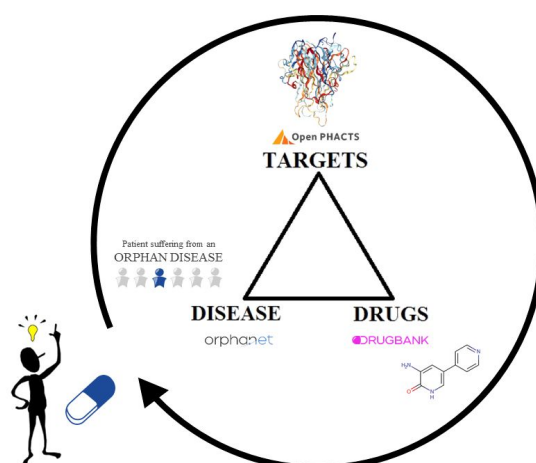


ABSTRACT

Triangulation of repurposing candidates for orphan diseases

JANA GURINOVA, *University of Vienna*

GRAPHICAL ABSTRACT



An orphan disease is by definition a disease that affects very few patients, the threshold in the US being less than 200,000 patients. Some orphan diseases are quite well known, such as cystic fibrosis, and some are virtually unknown, especially if they affect only a handful of patients worldwide. The diseases are rare, the patients however are numerous. Up to 8,000 orphan diseases have been described to date, leading to estimations of roughly 25 million patients suffering from rare diseases in the US and around 400 million worldwide.¹ This is in stark contrast to the 496 marketing approvals given to products for the treatment of orphan diseases in the US.²

Over the last decade there has been an increase in interest in therapeutics for orphan diseases due to their potential of greater profitability when considered across their lifetime (final average valuation at US\$818 million³) compared to non-orphan drugs.⁴ Simultaneously, rising R&D costs coupled with a world-wide pressure on prices, challenges from generics and ever-increasing regulatory hurdles have given rise to the trend of drug repurposing⁵, the application of approved drugs for the treatment of new

¹ Kaplan W et al., Priority Medicines for Europe and the World 2013 Update,

http://www.who.int/medicines/areas/priority_medicines/MasterDocJune28_FINAL_Web.pdf, WHO, July 2013

² To arrive at the number of drugs with at least one marketing approval for an orphan disease the dataset from

<https://www.accessdata.fda.gov/scripts/opdlisting/oopd/index.cfm> was downloaded and analyzed in KNIME.

³ Fagnan D. E. et al., Financing Drug Discovery for Orphan Diseases, *Drug Discovery Today*, May 2014

⁴ The Economic Power of Orphan Drugs, *Thomson Reuters*, 2012

⁵ Ted T. Ashburn et al., Drug repositioning: identifying and developing new uses for existing drugs, *Nature Reviews Drug Discovery*, August 2004

JANA GURINOVA
University of Vienna, Austria
jana.gurinova@univie.ac.at
ID: CINF 66 (Control ID 2869475)



indications. A significant advantage for the patient community lies in the fact that drug repurposing can provide safe drugs in a fraction of the time it would take to develop a new drug. Quite a few articles have been written over the last couple of years about drug repurposing. The majority focuses on drug repurposing for specific diseases or disease classes, while a small number elaborates on *in silico* methods for comprehensive repurposing across many diseases.⁶ Both the pharma industry as well as the rare disease patient community could obviously profit from a general repurposing method being applied to orphan diseases. The work described in this abstract is such a method, and to my knowledge it is the sole existing attempt at comprehensive drug repurposing for orphan diseases. It relies exclusively on data mining and results in an enrichment of useful associations between diseases and drugs of 58%. This means that in a manually inspected subset of results roughly two out of three associations between a disease and a drug (as a candidate for repurposing) were found to be either mentioned in publications, determined to be useful by a manual check or under investigation in clinical trials. Such a high enrichment indicates that the method is able to retrieve relevant associations. The method is a triangulation between diseases, their targets and drugs that act on these targets. Such a triangulation is able to retrieve previously unknown associations between diseases and drugs by expanding the connection between a disease and its drug by introducing an intermediate step where target information is taken into account. KNIME⁷ is the workflow management system that was used to execute this triangulation.

THE WORKFLOW

The workflow is based on the integration of three powerful databases. In a nutshell the workflow starts with a list of orphan diseases from the Orphanet⁸ database which is then connected to the Open PHACTS Discovery Platform⁹ for the retrieval of targets, and the targets are subsequently connected to respective drugs from DrugBank¹⁰. In this manner previously unknown associations between diseases and drugs can be retrieved. The idea behind this approach is that one target can be related to multiple diseases, and that a drug that has been developed for such a target in the context of one disease can potentially be used for the treatment of another disease. Additionally rare diseases are quite unknown and their targets therefore also, meaning that often it is simply not known that a rare disease shares a target with a common disease and that a treatment for a common disease might be very useful for the treatment of an orphan disease.

Orphanet, the first database used in the workflow, is a resource dedicated to information about orphan diseases, providing for example cross-references to other databases and information about disease-target relationships. Orphanet data can be downloaded from Orphadata.¹¹ Two of the provided datasets were used in the workflow, the “Rare diseases and cross-referencing” dataset and the “Rare diseases with their associated genes” dataset. The first was used because it provides multiple identifiers

⁶ Adam S. Brown and Chirag J. Patel, A review of validation strategies for computational drug repositioning, *Briefings in Bioinformatics*, November 2016

⁷ KNIME AG, KNIME Analytics Platform, <https://www.knime.com/>

⁸ Orphanet: an online database of rare diseases and orphan drugs, available at <http://www.orpha.net>, © INSERM, 1997

⁹ Alasdair J. G. Gray et al., Applying Linked Data Approaches to Pharmacology: Architectural Decisions and Implementation, *Semantic Web Journal*, April 2012

¹⁰ Vivian Law et al., DrugBank 4.0: shedding new light on drug metabolism, *Nucleic Acids Research*, January 2014

¹¹ Orphadata: Free access data from Orphanet, available at <http://www.orphadata.org>, © INSERM, 1997

JANA GURINOVA
 University of Vienna, Austria
 jana.gurinova@univie.ac.at
 ID: CINP 66 (Control ID 2869475)



(e.g. UMLS - Unified Medical Language System) for each of the listed diseases so that a disease from Orphanet can be connected to other databases, in this case to Open PHACTS for the retrieval of targets for orphan diseases. From the 9,557 diseases and their subtypes listed in the dataset 5,605 have a UMLS identifier which was used for the connection to Open PHACTS.

A disease in general is connected to a target in a certain way. This could be an up-regulation of the target (often in the case of cancers), a missense mutation (as can be the case for cystic fibrosis) or for example a duplication which leads to a deficiency (fragile X syndrome). The retrieval of targets for diseases is the second major step in the workflow and was done through Open PHACTS. Open PHACTS is a platform which combines various databases and allows for complex queries. One of them is DisGeNET¹², a database of genes associated with human diseases. 113,636 gene associations were retrieved for the list of rare diseases by accessing DisGeNET through Open PHACTS and 7,046 were present in the "Rare diseases with their associated genes" dataset from Orphadata. Some associations were present in both datasets which is why the final number of unique disease-target associations amounted to 117,851. Targets can be involved in multiple diseases which is why the number of unique targets from these disease-target associations amounted to 11,171, approximately double the number of diseases the workflow starts out with.

As the third step in the workflow drugs for these targets were retrieved from DrugBank, a comprehensive online database which contains information on drugs and drug targets. The drugs were then filtered for those which are approved (viable for repurposing instead of e.g. withdrawn). The information was at first retrieved from DrugBank through Open PHACTS and enriched with the 'action' annotations (e.g. drug is an inhibitor) directly from DrugBank. This step resulted in 3,664 unique associations between targets and drugs which all in all resulted in 65,303 unique disease-drug associations for possible repurposing.

As is to be expected not all of the associations are viable repurposing candidates. This is on the one hand due to the nature of text mining, which DisGeNET partially relies on, resulting for example in TNF antibodies as repurposing candidates for diseases which are in fact not caused by TNF directly, simply because upregulation of TNF is often mentioned in publications as an accompanying symptom and TNF is therefore retrieved as an association for the disease by text mining algorithms, and on the other hand some associations are not viable because of the difficulty to determine the type of connection between a disease and its target and therefore the necessary drug type. For example if the disease is caused by an upregulation of a certain target, the drug needs to be an inhibitor. The information on the type of connection between a disease and its target however is not present in any database and can only be determined by a manual curation of the results.

To streamline the curation I began to implement a text mining approach that mimics a manual curation. The idea is to analyze abstracts as a researcher would when attempting to determine the type of association between a disease and a target. Usually the abstract of the respective publication provides sufficient information to determine the type of association (e.g. upregulation). Viable drug associations can be curated with ease once the type of association is determined. The text mining approach therefore analyses abstracts and automatically assigns the type of association. As with any predictive method an excellent training set is essential and even then every method has its limitations. Text mining

¹² Janet Piñero et al., DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic Acids Research*, October 2016

JANA GURINOVA
University of Vienna, Austria
jana.gurinova@univie.ac.at
ID: CINP 66 (Control ID 2869475)



in KNIME is computationally expensive and therefore quite time-consuming, and as a method generally handles explicit and short textual entries better than scientific publications. My current training set does not yield satisfactory results and therefore this part of the workflow is currently under construction.

VALIDATION

Two validations were applied to measure the performance of the workflow. In the field of *in silico* repurposing three types of validation are common, a validation with a single example or case study of a single disease area (CSV), a sensitivity-based validation only (SV) and a sensitivity- and specificity-based validation (SSV).¹³

An SV was done first for this method using the Rare Disease Repurposing Database (RDRD)¹⁴ from the FDA, which unfortunately is not available anymore. The RDRD listed products that had received orphan status designation, meaning that they had been found to be 'promising' for the treatment of a rare disease. A cross-check between the results of the workflow and the products listed in the RDRD was performed for the validation, resulting in a subset of 205 disease-drug associations. This subset was then analyzed manually to determine the quality of the workflow results. Detailed figures from the analysis can be seen in Figure 1.

In 119 (58.0%) cases the associations were determined to be useful. These were associations which were present in scientific publications, associations with a potentially beneficial triangulation, associations that were or had been under investigation in clinical trials and one association with an approval by the FDA. In some cases the usefulness could not be determined and roughly 28% of the associations were determined not to be useful. All in all roughly 60% of the results were determined to be viable for repurposing, constituting a very good enrichment.

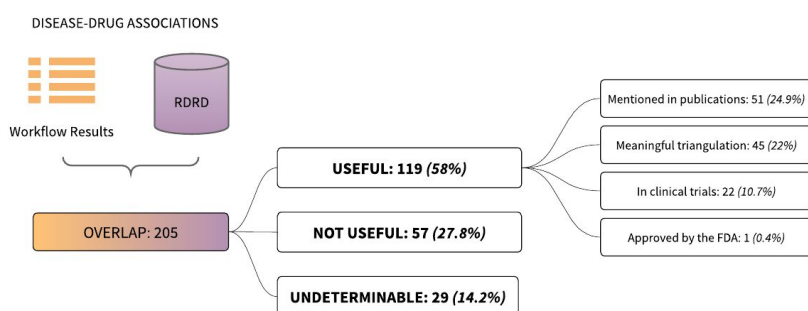


Figure 1: Results from the sensitivity-based validation

¹³ Adam S. Brown and Chirag J. P., A review of validation strategies for computational drug repositioning, *Briefings in Bioinformatics*, November 2016

¹⁴ Xu Kui et al., Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases, *Briefings in Bioinformatics*, February 2011

JANA GURINOVA
University of Vienna, Austria
jana.gurinova@univie.ac.at
ID: CINP 66 (Control ID 2869475)



The second validation of this method was an SSV and was performed with the recently introduced database for drug repositioning called repoDB by Adam S. Brown and Chirag J. Patel.¹⁵ RepoDB is a gold standard database for drug repositioning that consists of both true positives (approved drug-indication pairs) and true negatives (failed drug-indication pairs) and is meant to unify and facilitate a comparison of *in silico* drug repurposing methods. Approved indications in the repoDB originate from DrugCentral¹⁶ and failed indications originate from the Aggregate Analysis of ClinicalTrials.gov Database¹⁷. A cross-check of the results from the workflow and repoDB resulted in a subset of 331 associations which were then analyzed further. Detailed figures from the analysis can be seen in Figure 2.

Of the 331 overlapping disease-drug associations 60.4% were approved, showing an enrichment comparable to the SV validation (58.0%). Roughly 40% of the overlapping associations failed in their attempt at repurposing during clinical trials, being either suspended, terminated or withdrawn due to various reasons. Interestingly, of the 102 failed associations annotated with descriptions, the most common reason for failure was a lack of enrollment of patients. Only a handful of trials failed due to safety or efficacy concerns.

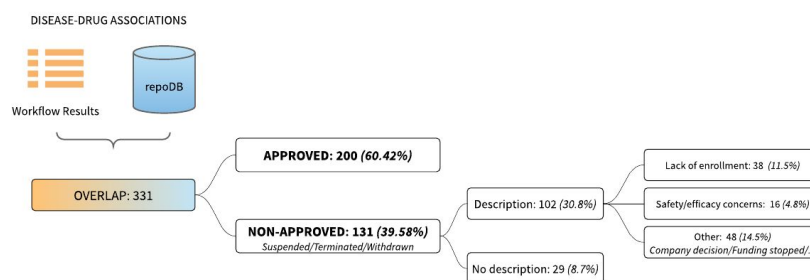


Figure 2: Results from the sensitivity- and specificity-based validation

Both validation methods yield matching results and suggest that roughly two out of three associations retrieved by the triangulation method are indeed viable for repurposing.

For some of the roughly 8000 orphan diseases, drug repurposing is the only hope of gaining access to approved therapeutics. The above mentioned method is a systematic approach that creates value for the patient community as well as for the pharma industry.

¹⁵ Brown A. S. and Chirag J.P., A standard database for drug repositioning, *Scientific Data*, March 2017

¹⁶ Ursu O. et al., DrugCentral: online drug compendium, *Nucleic Acids Research*, October 2016

¹⁷ Clinical Trials Transformation Initiative, Aggregate Analysis of ClinicalTrials.gov, <https://aact.ctti-clinicaltrials.org/>

BIBLIOGRAPHY

1. Kaplan, W., Wirtz, V. J., Mante-Teeuwisse, A., Stolk, P., Duthey, B. & Laining, R. *Priority Medicines for Europe and the World 2013 Update* Accessed on 09.02.2018. <http://www.who.int/medicines/areas/priority_medicines/MasterDocJune28_FINAL_Web.pdf> (2013).
2. Franco, P. Orphan drugs: the regulatory environment. *Drug Discovery Today* **18**, 163–172. ISSN: 1359-6446 (2013).
3. The European Parliament and the Council of the European Union. *REGULATION (EC) No 141/2000 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 1999 on orphan medicinal products* Accessed on 09.02.2018. <<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:018:0001:0005:en:PDF>> (2000).
4. Graf von der Schulenburg, J.-M. & Frank, M. Rare is frequent and frequent is costly: rare diseases as a challenge for health care systems. *The European Journal of Health Economics* **16**, 113–118. ISSN: 1618-7601 (2015).
5. European Commission. *Inventory of Union and Member State incentives to support research into, and the development and availability of, orphan medicinal products — state of play 2015* Accessed on 09.02.2018. <https://ec.europa.eu/health/sites/health/files/files/orphanmp/doc/orphan_inv_cwd_20160126.pdf> (2016).
6. *Demographics of the European Union Eurostat* Accessed on 31.10.2018. <<https://ec.europa.eu/eurostat/web/main/home>>.
7. *US Census Bureau* Accessed on 31.10.2018. <<http://www.census.gov/population/www/popclockus.html>>.
8. *Portal site of official Statistics of Japan* Accessed on 31.10.2018. <<http://www.e-stat.go.jp/SG1/estat/ListE.do?bid=000001032402&cyclo=0>>.
9. *Australian Bureau of Statistics* Accessed on 31.10.2018. <<http://www.abs.gov.au/>>.
10. *Statistics Canada* Accessed on 31.10.2018. <<https://www.statcan.gc.ca/eng/start>>.
11. *National Population and Talent Division* Accessed on 31.10.2018. <<https://www.gov.sg/sgdi/ministries/pmo/departments/sg/departments/nptd>>.
12. *National Statistics Republic of China (Taiwan)* Accessed on 31.10.2018. <<http://eng.stat.gov.tw/mp.asp?mp=5>>.
13. *Rare Disease and Orphan Drug Act (Twain)* Accessed on 31.10.2018. <<https://www.hpa.gov.tw/EngPages/Detail.aspx?nodeid=1058&pid=6031>>.
14. Fagnan, D. E., Gromatzky, A. A., Stein, R. M., Fernandez, J.-M. & Lo, A. W. Financing drug discovery for orphan diseases. *Drug Discovery Today* **19**, 533–538. ISSN: 1359-6446 (2014).

15. FDA. *Search engine for orphan drug designations and approvals* Accessed on 16.04.2018. <<https://www.accessdata.fda.gov/scripts/opdlisting/oopd/index.cfm>>.
16. Murakami, M. & Narukawa, M. Matched analysis on orphan drug designations and approvals: cross regional analysis in the United States, the European Union, and Japan. *Drug Discovery Today* **21**, 544–549. ISSN: 1359-6446 (2016).
17. Mullard, A. 2015 FDA drug approvals. *Nature Reviews Drug Discovery* **15**, 73–76. ISSN: 1474-1784 (2016).
18. Mullard, A. 2016 FDA drug approvals. *Nature Reviews Drug Discovery* **16**, 73–76. ISSN: 1474-1784 (2017).
19. Thomson Reuters. *The Economic Power of Orphan Drugs* Accessed on 31.10.2018. <<http://dtr-pharma.com/wp-content/uploads/2018/09/THE-ECONOMIC-POWER-of-Orphan-Drugs.pdf>> (2012).
20. Amgen. *Amgen 2011 Annual Report and Financial Summary* Accessed on 09.02.2018. <<http://phx.corporate-ir.net/External.File?item=UGFyZW50SUQ9MTM0MzEzEENoaWxkSUQ9LTF8VHlwZT0z&t=1>>.
21. Hill, A., Platts, P. J., Smith, A., Richards, S. J., Cullen, M. J., Hill, Q. A., Roman, E. & Hillmen, P. The Incidence and Prevalence of Paroxysmal Nocturnal Hemoglobinuria (PNH) and Survival of Patients in Yorkshire. *Blood* **108**, 985–985. ISSN: 0006-4971 (2015).
22. Herper, M. *The World's most expensive drugs* Accessed on 26.10.2017. <<https://www.forbes.com/2010/02/19/expensive-drugs-cost-business-healthcare-rare-diseases.html>>.
23. Borski, K. Economic aspects of rare diseases. *Developmental period medicine* **19**, 528–532. ISSN: 1428-345X (2015).
24. Angelis, A., Tordrup, D. & Kanavos, P. Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. *Health Policy* **119**, 964–979. ISSN: 0168-8510 (2015).
25. López-Bastida, J., Oliva-Moreno, J., Linertová, R. & Serrano-Aguilar, P. Social/economic costs and health-related quality of life in patients with rare diseases in Europe. *The European Journal of Health Economics* **17**, 1–5. ISSN: 1618-7601 (2016).
26. Muthyala, R. Orphan/rare drug discovery through drug repositioning. *Drug Discovery Today: Therapeutic Strategies* **8**. Drug repurposing, 71–76. ISSN: 1740-6773 (2011).
27. Sardana, D., Zhu, C., Zhang, M., Gudivada, R. C., Yang, L. & Jegga, A. G. Drug repositioning for orphan diseases. *Briefings in Bioinformatics* **12**, 346–356. ISSN: 1477-4054 (2011).
28. Volmar, C.-H., Wahlestedt, C. & Brothers, S. P. Orphan diseases: state of the drug discovery art. *Wiener Medizinische Wochenschrift* **167**, 197–204. ISSN: 1563-258X (2017).

29. Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jørgensen, H. L., Cox, I. J., Hansen, L. K., Ingwersen, P. & Winther, O. FindZebra: A search engine for rare diseases. *International Journal of Medical Informatics* **82**, 528–538. ISSN: 1386-5056 (2013).
30. Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S. & Robinson, P. N. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *The American Journal of Human Genetics* **85**, 457–464. ISSN: 0002-9297 (2009).
31. Müller, T., Jerrentrup, A. & Schäfer, J. R. Computerunterstützte Diagnosefindung bei seltenen Erkrankungen. *Der Internist*. ISSN: 1432-1289. doi:10.1007/s00108-017-0218-z (Mar. 2017).
32. Orphanet encyclopedia. *Orphanet: an online database of rare diseases and orphan drugs*. ©, INSERM 1997. Accessed on 23.10.2017. <<http://www.orphanet>>.
33. Orphadata. *Orphadata: Free access data from Orphanet*. ©, INSERM 1997 Accessed on 19.01.2018. <<http://www.orphadata.org/cgi-bin/index.php/>>.
34. Marusina, K., Welsch, D., Rose, L., Brock, D., Bahr, N., M. Cohen, A., A. Gacel-Sinclair, R., Vang, P., G. Ruminski, P., E. Bloom, B., Nagasawa, P. & P. Guo, B. in *Collaborative Innovation in Drug Discovery: Strategies for Public and Private Partnerships* (ed Chaguturu, R.) 115–133 (John Wiley & Sons, Inc., Apr. 2014). ISBN: 9780470917374.
35. Xu, K. & Côté, T. R. Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases. *Briefings in Bioinformatics* **12**, 341–345. ISSN: 1477-4054 (2011).
36. Ekins, S., Williams, A. J., Krasowski, M. D. & Freundlich, J. S. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discovery Today* **16**, 298–310. ISSN: 1359-6446 (2011).
37. Brown, A. S. & Patel, C. J. A review of validation strategies for computational drug repositioning. *Briefings in Bioinformatics* **19**, 174–177. ISSN: 1477-4054 (2018).
38. Orphadata. *Rare diseases and cross-referencing* Accessed on 31.10.2018. <http://www.orphadata.org/data/xml/en_product1.xml>.
39. Orphadata. *Rare diseases with their associated genes* Accessed on 31.10.2018. <http://www.orphadata.org/data/xml/en_product6.xml>.
40. NIH. *Unified Medical Language System* Accessed on 19.01.2018. <<https://www.nlm.nih.gov/research/umls/>>.
41. Open PHACTS Foundation. *Open PHACTS Discovery Platform* Accessed on 19.10.2017. <<https://www.openphacts.org/>>.
43. Queralt-Rosinach, N., Piñero, J., Bravo, A., Sanz, F. & Furlong, L. I. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. *Bioinformatics* **32**, 2236–2238 (2016).

44. Willighagen, E. L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A. J., Tkachenko, V., Hastings, J., Chen, B. & Wild, D. J. The ChEMBL database as linked open data. *Journal of Cheminformatics* **5**, 23. ISSN: 1758-2946 (2013).
45. Integrative Biomedical Informatics Group. *DisGeNET* Accessed on 23.10.2017. <<http://www.disgenet.org/>>.
46. Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F. & Furlong, L. I. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, bav028 (2015).
47. Piñero, J., Bravo, A., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F. & Furlong, L. I. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**, D833–D839. ISSN: 1362-4962 (2017).
48. Wishart Research Group. *DrugBank* Accessed on 23.10.2017. <<https://www.drugbank.ca/>>.
49. Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. & Woolsey, J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Research* **34**, D668–D672. ISSN: 1362-4962 (2006).
50. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082. ISSN: 1362-4962 (2018).
51. Brown, A. S. & Patel, C. J. A standard database for drug repositioning. *Scientific data* **4**, 170029 (2017).
52. EMBL-EBI. *EMBL-EBI web services* Accessed on 26.04.2018. <<https://www.ebi.ac.uk/seqdb/confluence/display/WEBSERVICES/EMBL-EBI+Web+Services>>.
53. *European Molecular Biology Laboratory* Accessed on 26.04.2018. <<https://www.embl.org/>>.
54. *European Bioinformatics Institute* Accessed on 26.04.2018. <<https://www.ebi.ac.uk/>>.
55. McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P. & Lopez, R. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research* **41**, W597–W600 (2013).
56. Lopez, R., Cowley, A., Li, W. & McWilliam, H. Using EMBL-EBI Services via Web Interface and Programmatically via Web Services. *Current protocols in bioinformatics*, 3–12 (2014).
57. Europe PMC. *Articles RESTful API - RESTful Web Service* Accessed on 26.04.2018. <<https://europepmc.org/RestfulWebService>>.
58. Europe PMC Funders' Group. *Europe PubMed Central* Accessed on 26.04.2018. <<https://europepmc.org/>>.

59. National Center of Biotechnology Information. *PubMed* Accessed on 26.04.2018. <<https://www.ncbi.nlm.nih.gov/pubmed>>.
60. National Center of Biotechnology Information. *PubMed Central*[®] Accessed on 26.04.2018. <<https://www.ncbi.nlm.nih.gov/pmc/>>.
61. Tomasso, P. D. *Awesome Pipeline* Accessed on 27.04.2018. <<https://github.com/pditommaso/awesome-pipeline>>.
62. Crusoe, M. R. *Computational Data Analysis Workflow Systems* Accessed on 27.04.2018. <<https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems>>.
63. KNIME AG. *KNIME*[®] *Analytics Platform* Accessed on 27.04.2018. <<https://www.knime.com/knime-analytics-platform>>.
64. BIOVIA. *Pipeline Pilot* Accessed on 27.04.2018. <<http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/>>.
65. Warr, W. A. Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design* **26**, 801–804. ISSN: 1573-4951 (2012).
66. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K. & Wiswedel, B. KNIME - the Konstanz Information Miner: Version 2.0 and Beyond. *SIGKDD Explor. Newsl.* **11**, 26–31. ISSN: 1931-0145 (2009).
67. Meinl, T., Wiswedel, B. & Berthold, M. R. in *Computational Approaches in Cheminformatics and Bioinformatics* (eds Guha, R. & Bender, A.) 182–184 (John Wiley & Sons, 2012). ISBN: 978-1-118-13142-8.
68. KNIME AG. *Getting Started* Accessed on 27.04.2018. <<https://www.knime.com/knime>>.
69. KNIME GmbH. *KNIME Quickstart Guide* Accessed on 27.04.2018. <https://files.knime.com/sites/default/files/inline-images/KNIME_quickstart.pdf>.
70. KNIME AG. *Example Workflows* Accessed on 27.04.2018. <<https://www.knime.com/example-workflows>>.
71. KNIME[®]. *Basic XML processing with XPath* Accessed on 27.04.2018. <<https://www.knime.com/nodeguide/data-access/structured-data/basic-xml-processing-with-xpath>>.
72. Open PHACTS Foundation. *OpenPHACTS Development Version API Active Docs* Accessed on 29.01.2018. <<https://dev.openphacts.org/docs/2.2>>.
73. Digles, D., Caracoti, A. & Jacoby, E. in *Phenotypic Screening: Methods and Protocols* (ed Wagner, B.) 183–193 (Springer New York, New York, NY, 2018). ISBN: 978-1-4939-7847-2. doi:10.1007/978-1-4939-7847-2_14. <https://doi.org/10.1007/978-1-4939-7847-2_14>.
74. KNIME AG. *Text Processing Applications* Accessed on 05.05.2018. <<https://www.slideshare.net/KNIMESlides/text-processing-with-knime>>.

75. KNIME AG. *Text Mining Webinar* Accessed on 05.05.018. <https://files.knime.com/sites/default/files/inline-images/slides_textminingwebinar_20131030.pdf>.
76. KNIME AG. *Text Mining Webinar Video*, accessed on 05.05.2018. <<https://www.youtube.com/watch?v=rY7vpTLYllg>>.
77. Dormer, R. L., Harris, C. M., Clark, Z., Pereira, M. M. C., Doull, I. J. M., Norez, C., Becq, F. & McPherson, M. A. Sildenafil (Viagra) corrects $\Delta F508$ -CFTR location in nasal epithelial cells from patients with cystic fibrosis. *Thorax* **60**, 55–59. ISSN: 0040-6376 (2005).
78. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**, 3191–3192 (2005).
79. Sinha, S., Mishra, S. K., Sharma, S., Patibandla, P. K., Mallick, P. K., Sharma, S. K., Mohanty, S., Pati, S. S., Mishra, S. K., Ramteke, B. K., Bhatt, R., Joshi, H., Dash, A. P., Ahuja, R. C., Awasthi, S., Indian Genome Variation Consortium, Venkatesh, V. & Habib, S. Polymorphisms of TNF-enhancer and gene for Fc γ RIIa correlate with the severity of falciparum malaria in the ethnically diverse Indian population. *Malaria Journal* **7**, 13. ISSN: 1475-2875 (2008).
80. Emond, M. J., Louie, T., Emerson, J., Zhao, W., Mathias, R. A., Knowles, M. R., Wright, F. A., Rieder, M. J., Tabor, H. K., Nickerson, D. A. *et al.* Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nature genetics* **44**, 886 (2012).
81. Emond, M. J., Louie, T., Emerson, J., Chong, J. X., Mathias, R. A., Knowles, M. R., Rieder, M. J., Tabor, H. K., Nickerson, D. A., Barnes, K. C. *et al.* Exome sequencing of phenotypic extremes identifies CAV2 and TMC6 as interacting modifiers of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *PLoS genetics* **11**, e1005273 (2015).
85. American Chemical Society. *255th ACS National Meeting & Exposition* Accessed on 23.10.2017. <<https://global.acs.org/events/255th-ac-s-national-meeting-exposition/>>.